Relationship of fecal indicator bacteria to demographic patterns and land use in Austin, Texas
SR-13-10, May 2013

Abel Porras, P.E., Aaron Richter, E.I.T., Chris Herrington, P.E.

City of Austin
Watershed Protection Department
Environmental Resource Management Division

Abstract

The City of Austin assessed the relationship of existing Escherichia coli *fecal indicator bacterial data from non-storm influenced stream samples collected during the 2006-2012 Environmental Integrity Index (EII) sampling campaigns with associated demographic and land use data. The three analytical methods were used including a geostatistical analysis, a multiple linear regression analysis and a Random Forest model. Geostatistical analysis suggests higher density of median non-storm* E. coli *in the urban core and northern (or more developed) portions of the eastern watersheds relative to the western watersheds. By multiple linear regression analysis, flood detention structural controls were inversely related to bacteria counts while impervious cover in the riparian zone, transportation land use, industrial land use, presence of treated wastewater discharges and household density were directly related to bacteria counts. Different explanatory variables emerge in multiple linear regression models when sampling reaches are separated into east, west and urban groupings.*

Introduction

The City of Austin routinely collects *Escherichia coli* fecal indicator bacteria data as part of the citywide Environmental Integrity Index (EII) sampling program. A portion of the EII is submitted via the Texas Clean Rivers Program to the Texas Commission on Environmental Quality (TCEQ) for evaluation pursuant to section 303(d) of the federal Clean Water Act. TCEQ (2012) has identified 4 Austin waterbodies that do not support their designated primary contact recreation use based on elevated levels of fecal indicator bacteria. A Total Maximum Daily Load (TMDL) allocation was initiated by TCEQ in 2013 to address the contact recreation use impairment of the 4 watersheds in Austin (Waller Creek upstream of MLK, Walnut Creek, Taylor Slough South, Spicewood/Foster Branch Tributary). Stakeholders in the affected watersheds including the City of Austin are participating in the development of an Implementation Plan to reduce fecal loading. At the request of the stakeholder Implementation Plan Coordinating Committee, the City of Austin reviewed existing recent *E. coli* data to determine if relationships existed between bacteria concentrations and demographic or land cover information in the intervening drainage areas to the monitoring locations. This analysis may guide the selection or prioritization of fecal load reduction strategies in the Implementation Plan by indicating correlations with potential sources of fecal bacteria in Austin streams.

Input Data

Bacterial Data

E. coli counts were obtained from the City of Austin EII program for all monitoring reaches sampled at least three times from 2006 to 2012. The EII program intentionally targets flowing, non-storm influenced conditions across a wide gradient of human disturbance with quarterly grab samples collected every other year (City of Austin, 2002). There were 111 EII reaches with sufficient data. Plots of the raw data, broken down by watershed and geographic region, are included in Appendix A. The regions are partitioned into the eastern watersheds generally east of Interstate Highway 35, the urban watersheds in downtown central Austin, and the generally western watersheds west of Interstate Highway 35. The Austin urban core was defined as watersheds that roughly fell in between I-35, Mopac, SH-71, and US-183. This limited the watersheds to Shoal, Waller, Harper's Branch, Johnson, the upper reaches of Little Walnut and Buttermilk Creeks.

Generally, the eastern and western watersheds had median bacteria counts that were below 250 most probable number of *E. coli* colonies (MPN) per 100 mL. The Buttermilk Creek, Boggy Creek, Fort Branch and Harris Branch watersheds in the eastern watersheds and the Taylor Slough, West Bull Creek, and Walnut Creek watersheds in the western watersheds yielded higher median bacteria counts. Thus, any statistical model would look to see which factors follow this trend to explain the high bacteria counts for these watersheds. In contrast to the relatively low bacteria counts in the eastern and western watersheds, all of the urban watersheds consisted of median bacteria counts above 250 MPN/dL.

The number of samples from each sampling reach ranged from 6 to 14, with the variation in number resulting from variable drought conditions (only flowing, non-storm samples are collected for EII). Descriptive statistics including mean, median and geometric mean, were calculated for each watershed reach (Table B1 of Appendix B). Note that for many of these watershed reaches, the median of the data is a large distance from the mean of the data. This is an informal indication of skewness and suggests non-normality of the data. The bacteria counts were transformed by taking the fourth root of the counts (Table B2 of Appendix B), which has been recommended for use in benthic sampling programs (Downing, 1979) and may stabilize variance (Downing, 1981). The median and mean of these transformed data appear closer together, and thus was considered closer to a normal distribution. To further ameliorate the influence of outliers, the median of the transformed data was used in the statistical analysis presented.

Geographic Data

Geographic data was compiled for each sampling reach to examine the potential explanatory factors related to the variation in fecal indicator bacteria counts. The geographic data, extracted from City of Austin Geographic Information System databases, consisted of four categories (Table 1): demographics, infrastructure, land use, and land condition. Geographic information for each sampling reach was compiled for the contributing drainage area to the reach upstream to the adjacent downstream boundary of the next sampling reach. Thus, geographic information was effectively local and not reflective of the cumulative total upstream contributing drainage area. Number of households and population numbers were derived from Capital Area Metropolitan Planning Organization year 2007 estimates. Land use data was derived from the City of Austin 2006 classification. Vegetative cover information within 400 feet of the stream centerline was derived from the City of Austin (Scoggins et al., 2013) Index of Riparian Integrity.

Table 1: List of Factors used in the Analysis

Demographic and/or General Geographic Data	Number of Households in 2007 Population in 2007 Drainage.Area Location (2, if Urban,1 if East, 0 if West)
Data on Infrastructure	Weighted Average Diameter of Wastewater Lines Number of Permitted Wastewater Discharge Outfalls Number of Permitted Wastewater Land Application Facilities (>5,000 gallons per day) Total Length of Wastewater Collection Lines Number of Wastewater Lift Stations Number of On-Site Sewage Facilities (<5,000 gallons per day) Number of Flood Detention Ponds Number of Water Quality Ponds
Land Use Data (% of area)	Single Family (LU_100) Multi-Family (LU_200) Commercial (LU_300) Office Space (LU_400) Industrial (LU_500) Misc. Institutional Facilities (LU_600) Green Space (LU_700) Facilities related to Transportation (LU_800) Undeveloped (LU_900)
Land Condition Data (% of area within 400 feet of the stream centerline)	Robust Grasses (IRI_0) Sparse Grasses (IRI_1) Bare Ground (IRI_4) Woody Vegetation (IRI_5) Impervious Cover (IRI_6)

Methods of Statistical Analyses

Three types of statistical analyses were performed on the data: geostatistical analyses, multiple linear regression analyses, and Random Forest (Breiman, 2001) analyses. The following sections will briefly discuss the methods of these analyses and look at results from these analyses on the data.

Geostatistical Analyses

Geostatistical analysis is a common technique used in the natural sciences to predict a quantity at an unsampled location based on spatial continuity and its proximity to sampled locations (Isaaks and Srivastava, 1989). This technique derives from spatial autocorrelation principles (Tobler, 1970) and comes from the intuitive idea that things close together will be more similar than things farther apart. Geostatistics uses statistics and correlations between data pairs to get more information. In this case, the additional information is the variability in the data. Thus, geostatistics not only provides a prediction map of the quantity being sampled at every point on a map, but also an uncertainty map at every point on the map. This is helpful in providing a range of likely values for the quantity being measured. This technique is often called *kriging* after the South African mining engineer who developed this idea, Danie Krige.

Multiple Linear Regression Analysis
Multiple linear regression is used when a variety of factors are thought to influence a variable or quantity of interest (Sokal and Rohlf, 1995; Montgomery, Peck, and Vining, 2006). The regression can effectively determine which of the multiple factors are significant in explaining the quantity of interest. In this case, median bacteria counts are the quantity of interest and the multiple factors thought to influence this are the geographic data sets (Table 1).

Two approaches were used to determine examine the data. The first combined all the data together in one set. The second approach separated the data into their geographic locations. This stratification of the data is a common way of minimizing the variance in the data based on known physical differences. In this case, watersheds east of IH-35 were isolated because these watersheds have more gentle slopes and different soil conditions typical of the Blackland Prairie ecoregion. The urban core was separated from the rest due to the high impervious cover and long history of intensive human occupation. The remaining watersheds are all west of IH-35 (exclusive of the urban core) and include the high gradient streams and rocky soils of the Edwards Plateau ecoregion.

Random Forest Analysis
Random Forest analysis is an ensemble learning method for regression which creates multiple decision trees from random subsets of the data (Breiman, 2001). The final model can make predictions within the range of the training data set (a subset of the input data). Model error and variable importance are easily calculated from random forest output (Breiman, 2001).

Initially the data splits based on parameters of the regression model. This split happens at a node. Data continues to split at further nodes based on different parameter values. The process looks similar to the branching of tree limbs, thus it is called a decision tree. At each node the data will become more homogeneous (Gromping, 2009). It is common to grow the decision tree until there are only a few data points in the final decision. In some decision tree analysis the tree is then pruned, or the number of nodes is decreased based on an algorithm. Random forest analysis leaves the decision tree fully grown (Breiman, 2001).

Multiple versions of this decision tree analysis are run using different random subsets of the data. Predictions, error, and variable importance of the model are based on modes from all the decision trees. This method is nonparametric and does not require interactions to be specifically modeled as in linear regression models (Gromping, 2009). The interaction and nonlinearities in the data are "learned" by the model in the training data.

It has been suggested that random forest models better examine the contribution of each parameter in a multiple regression model over simpler models like simple linear regression or mixed effect regression (Strobl et al., 2009). In addition, random forest analysis is capable of handling cases where the number of predictors is large and the number of data points available for analysis is small (Breiman, 2001; Ishwaran, 2007; Strobl et al., 2007).

Four separate random forest analyses for the prediction of *E. coli* in the Austin watersheds were used. Watersheds were separated into east, west and urban groups as done in the multiple linear regression analysis. East and west watersheds were separated generally along the IH-35 corridor which mimics the ecoregion change between the Blackland Prairie and Edward's Plateau regions. Urban watersheds in this analysis match the urban watersheds specified in the City of Austin Land Development Code and include Shoal, Little Walnut, Waller, Johnson, Buttermilk, Boggy, Tannehill, Fort Branch, West Bouldin, East Bouldin, Blunn, and Harper's Branch (COA 2013).

To compensate for the difference in drainage acreage, parameters including length of wastewater lines, number of lift stations, number of OSSFs, number of TLAP permits, number of households, actual population, number of wastewater discharge permits, number of flood detention structural control ponds, and number of water quality structural control ponds were standardized for area. Other parameters were either percentages of area already or were not subject to change by drainage size. Random forest analysis can consist of a large number of trees to form accurate models (Strobl et al., 2009). It has been suggested that a large number of trees, as many as 1,000 trees, is particularly important when interested in the variable importance of a model (Breiman, 2002). This analysis used 600 trees for each random forest analysis and set the training data sets to 90% of the available data. Error estimation was then done on a random set of 10% of the data. Unlike the multiple linear regression analysis which used the median of the 4th root of the *E. coli* data, the medians of the untransformed *E. coli* counts by sampling reach were used in the Random Forest Analysis.

Results
<u>Geostatistical Analysis</u>
For the purposes of this report, a geostatistical analysis was performed as a preliminary examination of spatial patterns in bacteria counts and as a visual aid. A *kriged* map of the median bacteria counts from the stream reaches was generated (Figure 1). The redder colors represent higher median counts whereas the bluer colors represent lower median bacteria counts. The kriged surface is a representation of predicted values of bacteria counts in surface water under non-storm influenced conditions. The Austin urban core yielded high bacteria concentrations. Watersheds east of IH-35 appear to have higher bacteria counts than the western area, but the western watersheds had much higher variability, partially due to the inclusion of the upper reaches of Walnut Creek.

A "hot spot" in the eastern watershed group east of IH-35 at the upstream reach of the Harris Branch watershed is evident. Harris Branch receives treated wastewater effluent discharge, as does Gilleland Creek, although bacteria levels should be less than 126 MPN/dL based on TCEQ wastewater discharge permit conditions. To explore factors contributing to the observed variation in bacteria medians, multiple linear regression was used.
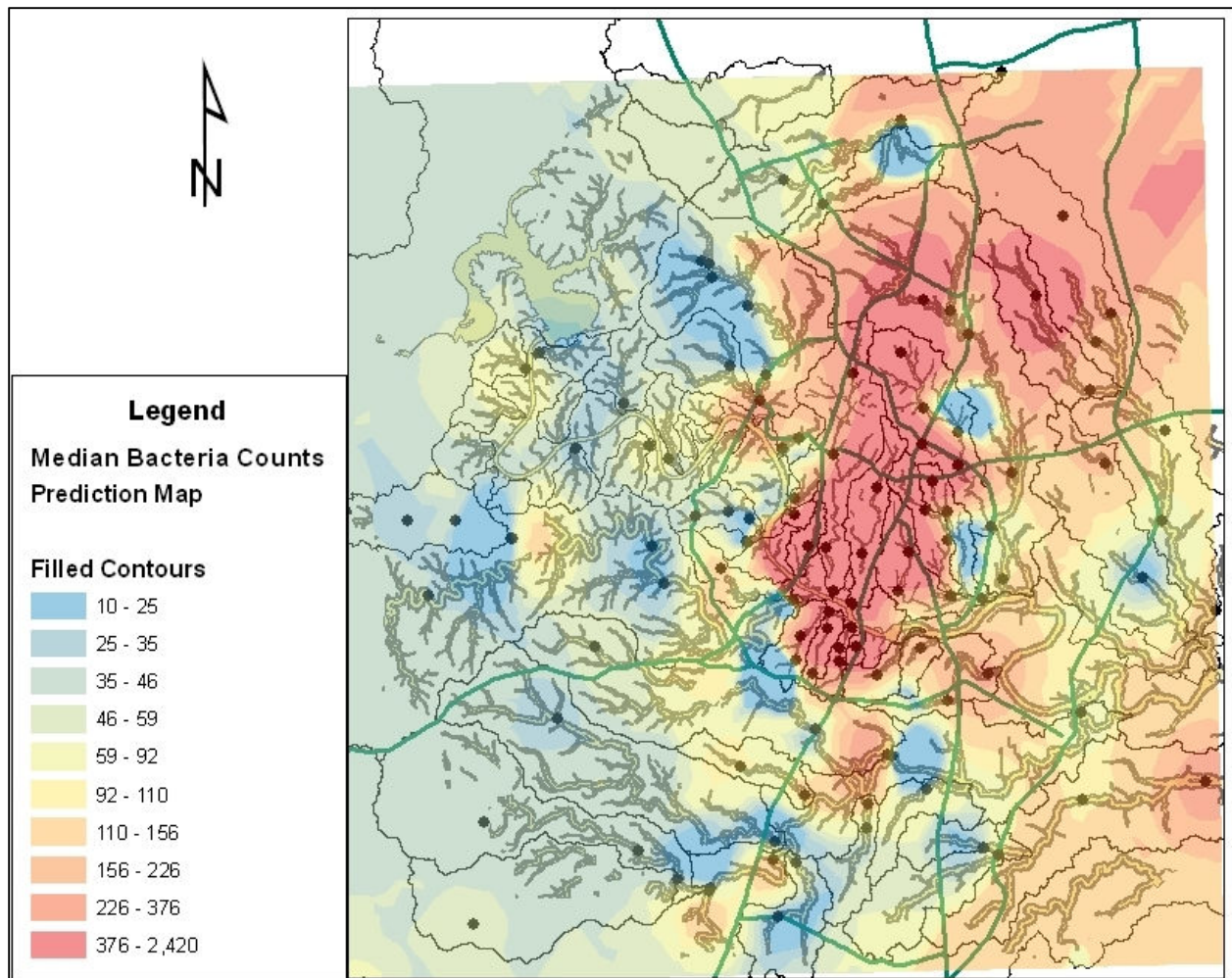
Figure 1: Kriged Map of Median Bacteria Counts in Austin, Texas.

Multiple Linear Regression
The first multiple linear regression approach took all the watersheds combined into one analysis and determined which of the multiple factors influence bacteria counts. Results from this analysis showed that, in general, the number of structural flood detention controls (ponds), number of households, the percentage of transportation facilities, the percentage of impervious cover in the riparian zone, the number of permitted wastewater discharge permits, and the percentage of industrial land use all significantly influence the median bacteria counts. The equation used to predict median bacteria counts for this data is:

$$y^{1/4} = -0.004 \cdot Ponds + 9 \cdot 10^{-5} Households + 4.17 \cdot \%Trans.Facility + 2.12 \cdot \%I.C.$$
$$+0.43 \cdot Discharge\ Permits + 2.5 \cdot \%Indust.Facility$$

This model had an adjusted R^2 value of 0.633 and an Akaike Information Criteria (AIC) score of -73.82. Thus, this model was a fair predictor of bacteria counts. However, it is a broad assessment of the bacteria in the watersheds of Austin. A finer resolution of the data is needed to isolate the different potential factors in the different ecosystems.

A second multiple linear regression approach stratified the watersheds into three respective geographic locations (east, west, urban core). Results from the stratified multiple linear regression analyses show

three different models (or equations) linking various factors with levels of bacteria counts for each of the watershed reaches.

First, under the urban watersheds (number of sampling reaches = 20), the median bacteria count was significantly influenced by the length of wastewater main lines, the number of structural flood detention controls (ponds), the percentage of single family land use, the percentage of institutional facility land use, the percentage of undeveloped land, the percentage of sparse grasses in the riparian zone, and the percentage of woody vegetation in the riparian zone. The Urban multiple linear regression watershed model is:

$$y_{URBAN}{}^{1/4} = 4.8 + 3.19 \cdot 10^{-6} \cdot WWMainLength - .015 \cdot Ponds - 4.9 \cdot \%SingleFam - 2.64 \cdot \%Inst.Facility - 18.9 \cdot \%Undeveloped + 14.7 \cdot \%SparseGrasses + 3.9 \cdot \%WoodyVegeation$$

The coefficient for the percentage of undeveloped land, single family land use and institutional facility land use is negative implying that as the percentage of these land uses increase, the median bacteria count is predicted to decrease. This is in contrast to the positive coefficients for the percentage of sparse grasses and unexpectedly the percentage of woody vegetation. Thus, this model appears to have some inherent contradictions.

This contradiction can be resolved in part by evaluating the coefficients as a weighting scheme. More weight is given by the model for the percentage of undeveloped land than for the percentage of sparse grasses and woody vegetation. Higher importance of undeveloped land is expected since undeveloped land is likely to have more infiltration and filtration of stormwater reducing bacteria loads.

The statistics for the Urban model selection show that the AIC number was -11.2 and the adjusted R^2 value was 0.72. The R^2 value is an indication of how much of the data variability is explained by the equation. In this case, 72% of the variability in the data is explained by the Urban model indicating that the model provided a good fit for data.

The eastern watershed multiple linear regression analysis contained 41 sampling reaches, but required only 2 variables to yield a model adequate to explain the data. For the eastern watersheds, median bacteria counts were influenced by the number of discharge permits and the household density:

$$y_{EASTERN}{}^{1/4} = 2.80 + 16,860 \cdot HouseholdDensity + 0.625 \cdot DischargePermits$$

It should be noted that three observations (or sampling reaches) were considered influential and had an overly large impact on the selection of the model variables. In this case, these observations occurred in the two Harris Branch reaches and the uppermost reach of Fort Branch. In the case of the Harris Branch reaches, the number of discharge permits was highly correlated to the higher median bacteria counts in that watershed although the wastewater effluent should be disinfected and have bacteria levels less than contact recreation standards. In the case of the Fort Branch watershed, the high household density was linked to the high median bacteria counts. These correlations overshadowed any other correlations from the other variables. Nevertheless, this simple model was relatively accurate in predicting the median bacteria count yielding an AIC number of -29.06 and an adjusted r^2 value of 0.43.

Removing these 3 influential sampling reaches changed the model so that the percentage of green space (permanently protected undeveloped land or park land) and undeveloped space were the significant factors although they are inversely related to bacteria counts. The resulting alternative Eastern model is:

$$y_{EASTERN}^{1/4} = 4.25 - 2.67 \cdot \%GreenSpace - 1.76 \cdot \%Undeveloped$$

The adjusted r^2 value of 0.30 for the alternative Eastern model is inadequate. The initial Eastern model appears to be a better predictor of median bacteria counts in the Eastern watersheds.

For the western watersheds (number of sampling reaches = 50), the multiple linear regression analysis provided this model:

$$y_{WESTERN}^{1/4} = 3.23 + 0.25 \cdot Length\ of\ WW - 6.32 \cdot \%Comm.Space + 9.04 \cdot \%IndustrialSpace - 3.98 \cdot \%GreenSpace + 5.05 \cdot \%Transp.Facilities - 16.08 \cdot \%BareGround - 1.36 \cdot \%WoodyVegetation$$

This model had an AIC of -61.4 and an adjusted r^2 value of 0.513. Note that for the western watersheds, increasing percentage of either bare ground within the riparian zone or the percentage of woody vegetation in the riparian zone *decreased* the median bacteria counts. This is in contrast to the Urban watershed models which indicated that median bacteria counts increased as woody vegetation increased. Bare ground in the riparian zone as derived from the Index of Riparian Integrity is different from and preferable to impervious cover in the riparian zone.

A difference between the urban and western watershed groups is that the population density in the western watersheds is considerably smaller than that of the urban watersheds. Thus, one possibility to explain this discrepancy is that the increased amount of human activity in the urban watersheds is influencing the increase in bacteria counts.

Validity of regression models must be determined by reviewing the diagnostic output. Multiple linear regression is based on the idea that the residuals (that is, the value predicted by the model subtracted from the actual value) are normally distributed with a mean of zero and a constant variance (Sokal and Rohlf, 1995). If these assumptions are not met, then other models must be attempted. To check these assumptions, the following three plots are utilized and presented in Appendix C for each of the models.

1. A plot of the fitted value to the residual is depicted to check the constant variance assumption. The plots for the urban and eastern watershed models fulfill these assumptions. The plot for the western watershed model shows some structure, but not enough to disqualify it.
2. A histogram of the residuals is depicted to show the normality of the residuals. The histograms for each of the watershed models show a slight deviation from a normal distribution.
3. A normal Q-Q plot of the residuals can be used to determine how far from normality the residuals can be given the number of observations. The dashed red lines in these plots provide limits for the residuals. The residuals for neither of the watershed models go beyond these limits.

The conformance of these plots to the assumptions shows that the models are a good fit to the data.

Random Forest Analysis
For all of the Austin watersheds (east, west and urban together), the random forest model predicted the use of the length of wastewater pipe, percentage of the riparian zone that was impervious cover, the density of households, population density, the percentage of land use that was transportation (streets and roads, land use code 800), and the percentage of undeveloped land use (land use code 900). The length of wastewater lines and the percent of the riparian zone that was impervious cover contributed most to the regression model (Figure 2). The overall percent of variation in *E. coli* variation explained by the model was 38.2%.
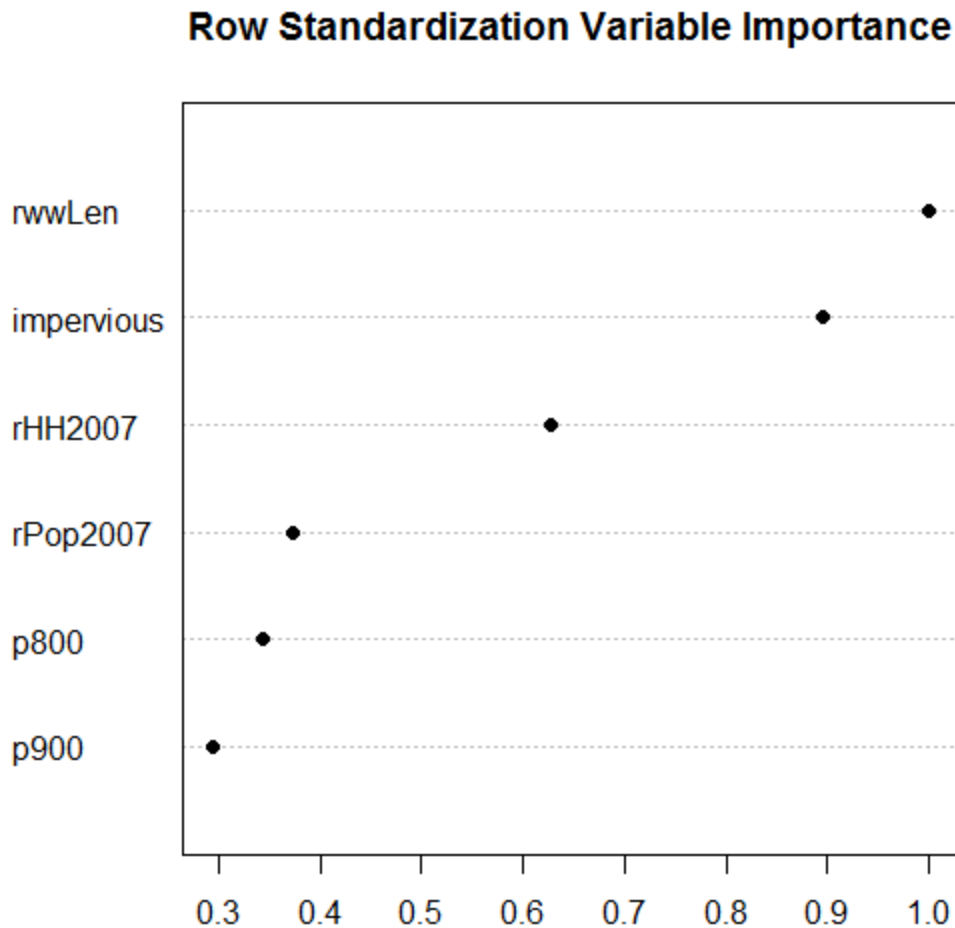
Row Standardization Variable Importance

Figure 4: Variable importance for the Random Forest model covering all Austin watersheds. Length of wastewater lines = rwwLen, impervious = impervious cover in the riparian zone, p800 = transportation land use, p900 = undeveloped land use, WW_Diameter = length-averaged diameter of wastewater lines, rHH2007 = year 2007 household density, rPop2007 = year 2007 population.

For the western watersheds separately, the random forest model predicted the use of the percentage of the riparian zone that was impervious cover, the percentage of the riparian zone that was woody vegetation, the percentage of land use that was industrial use (land use code 500), the density of households, the density of flood detention ponds, and the percentage of land use that was undeveloped (land use code 900). The percent of the riparian zone that was impervious cover and woody vegetation contributed most to the regression model (Figure 5). The overall percent of *E. coli* variation explained by the model was 50.05%.
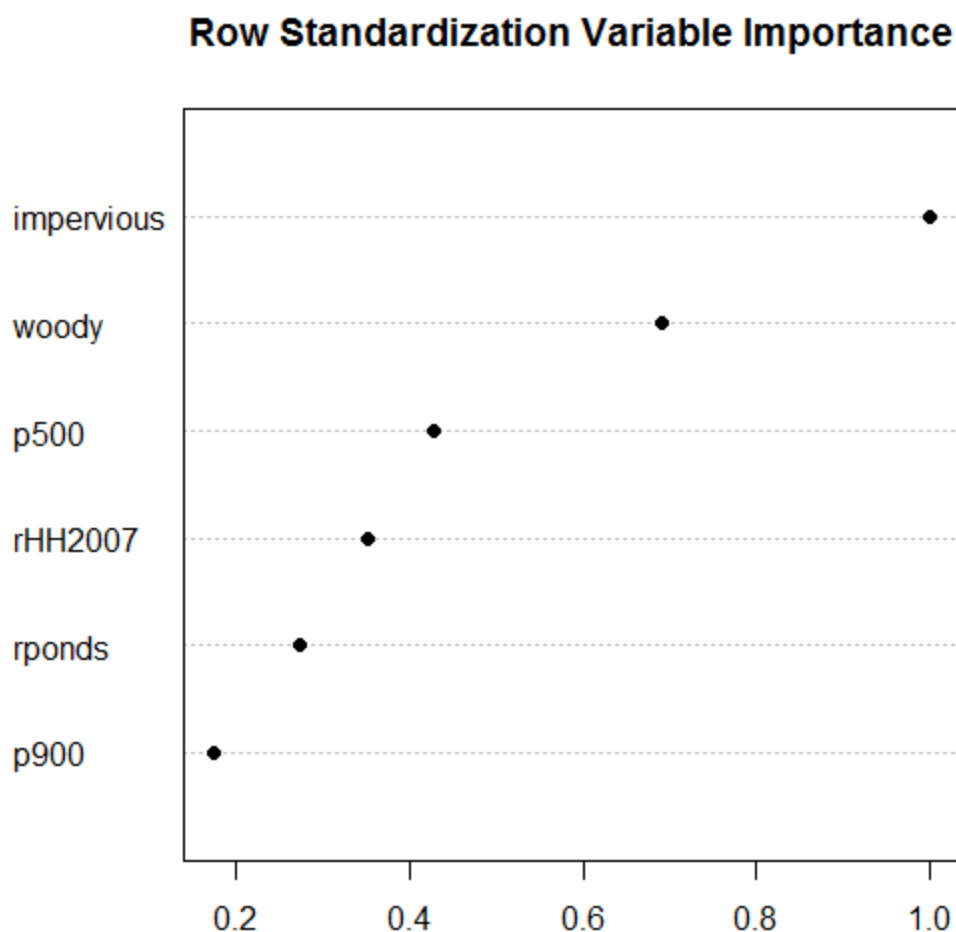
Row Standardization Variable Importance

Figure 5: Variable importance for the Random Forest model covering western Austin watersheds. Impervious = percent impervious cover in the riparian zone, woody = percent woody vegetation in the riparian zone, p500 = percent of industrial land use, rHH2007 = year 2007 density of households, rponds = density of flood detention ponds, p900 = percent of undeveloped land use.

For the eastern watersheds separately, the random forest model predicted the use of the percentage of land use that was commercial (land use code 300), the density of flood detention ponds, the population density, the percent of the riparian zone that was impervious cover, the density of water quality ponds, and the density of households in the reach. Each parameter suggested for use was important to the modeling (Figure 6). The overall percent of *E. coli* variation explained by the model was only 9.96%, suggesting low confidence in predictions made using this model.
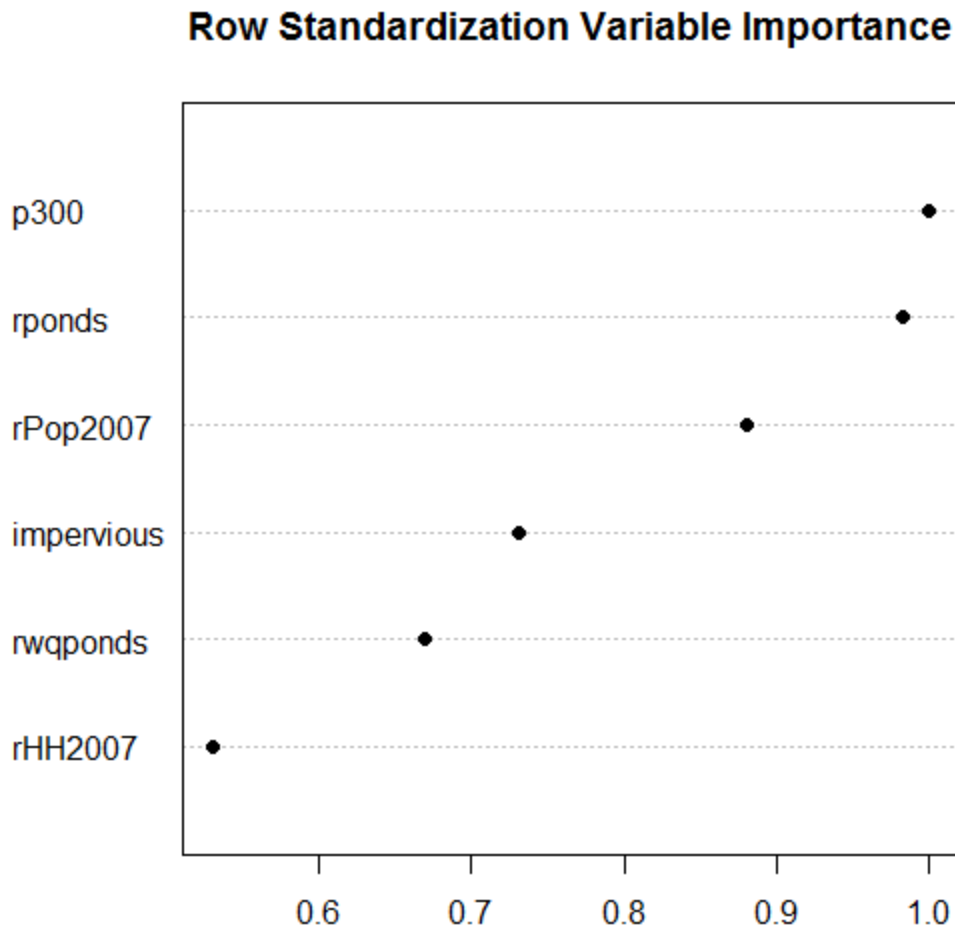
Row Standardization Variable Importance

Figure 6: Variable importance for the Random Forest model covering eastern Austin watersheds. P300 = commercial land use, rponds = density of flood detention ponds, rPop2007 = year 2007 population density, impervious = percent impervious cover in the riparian zone, rwqponds = density of water quality ponds, rHH2007 = year 2007 density of households.

For the urban watersheds separately, the random forest model predicted the use of the length of wastewater line in a reach, the percentage of transportation land use (land use code 800), the percentage of undeveloped land use (land use code 900), the length-average diameter of wastewater lines, the year 2007 density of households, and the percent of the riparian zone that was robust grasses. The length of the wastewater lines was the most important variable in the model (Figure 7). The overall percent of *E. coli* variation explained by the model was only 22.68%.
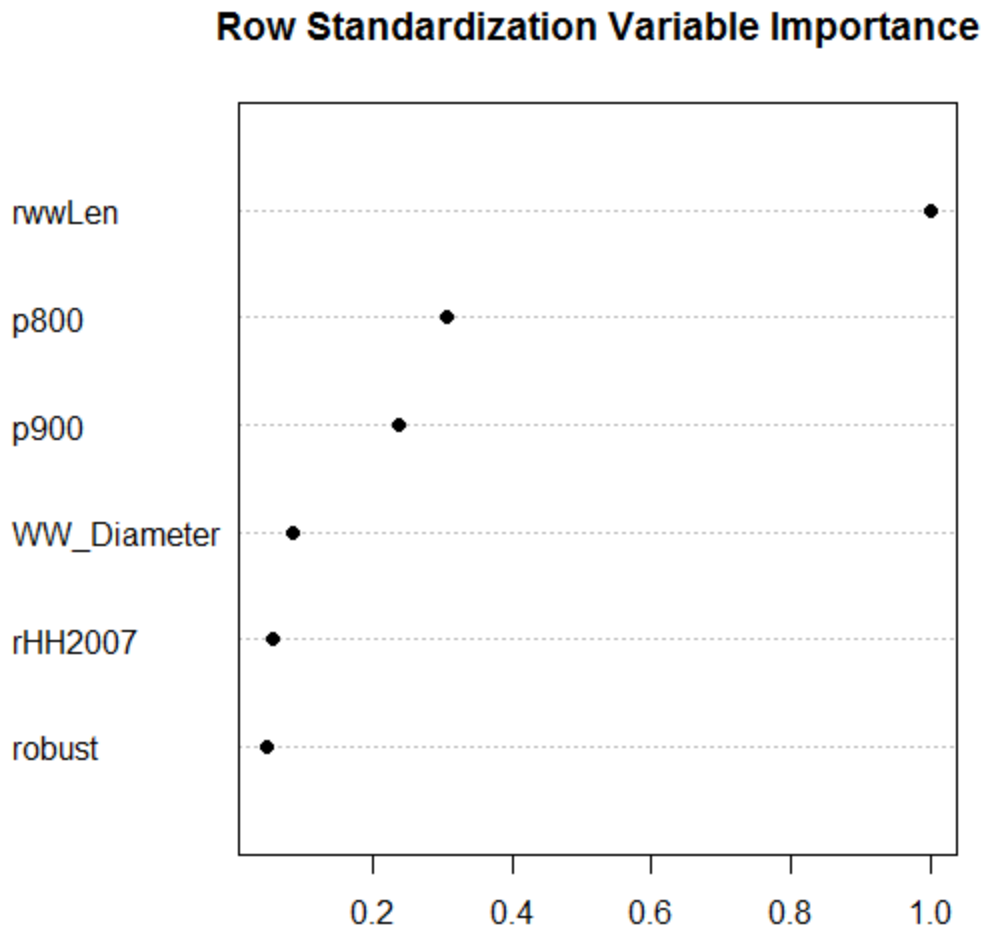
Row Standardization Variable Importance

Figure 7: Variable importance for the Random Forest model covering urban Austin watersheds. rwwLen = length of wastewater lines, p800 = percent transportation land use, p900 = percent undeveloped land use, WW_Diameter = length-averaged diameter of wastewater lines, rHH2007 = year 2007 density of households, robust = percent of riparian zone that was robust grasses.

Discussion

The kriged map of median bacteria counts (Figure 1) show higher density of median non-storm *E. coli* in the urban core and northern (or more developed) portions of the eastern watersheds relative to the western watersheds. Given the higher population density, older and larger wastewater infrastructure and reduced riparian zones, higher bacteria counts in urban areas are not unexpected. It is possible that the higher sediment loads in the deeper alluvial soils of the eastern watersheds influence the increased bacteria counts observed in Blackland Prairie Ecoregion streams assessed in this analysis. Sediments may act as reservoirs for fecal bacteria (Crabill et al., 1999; Sjogren, 1994), enhancing fecal bacteria survival (Craig et al., 2004) and reintroducing bacteria to the water column via resuspension of sediment (Stephenson 1982).

<u>Multiple Linear Regression</u>
The multiple linear regression models developed provide an examination of the factors which may relate to the observed variation in median bacteria counts in Austin watersheds. Citywide, flood detention structural controls were inversely related to bacteria counts while impervious cover in the riparian zone, transportation land use, industrial land use, presence of treated wastewater discharge permits and household density were positively related to bacteria counts.

When the data are separated by their geographic regions (urban, eastern, western), a different set of explanatory variables emerge. For the eastern watersheds, the number of treated wastewater discharge permits had a large influence on the high median bacteria counts particularly in the Harris Branch watersheds. The high household density influenced the high median bacteria counts particularly in the Fort Branch watersheds.

For the western watersheds, it appears that green space and woody vegetation decreases median bacteria counts, and industrial and transportation facilities (roads and parking lots) increases median bacteria counts. The inverse relationship in western watersheds between median bacteria and percentage of commercial facilities, opposite that of industrial facilities and roads, is not readily explained.

For the urban counts, the explanatory variables are more ambiguous. This regression model shows that undeveloped spaces, single family land use, flood detention structural controls and institutional land use decrease median bacteria counts, while wastewater collection line length, woody vegetation and sparse grasses in the riparian zone increase median bacteria counts. While this and other relationships between the geographic data and the bacteria counts may seem counter-intuitive, the model is merely picking variables that strongly influence the bacteria counts. It is possible that riparian areas in the urban watersheds are more intensively as an artifact of the higher population density in the urban areas. Classification categories within the Index of Riparian Integrity are mutually exclusive, and thus higher percentages of sparse grasses or woody vegetation may reflect unbalanced shifts in the three dimensional distribution (length, width and height) and diversity of riparian vegetation. Removal of understory vegetation beneath woody vegetation and mowing of grasses in riparian zones are frequent in managed urban parks and reduce riparian zone functionality (Richter and Duncan, 2012). While single family land use may be associated with some fecal load from domestic pet waste, land use categories are also mutually exclusive. Increased presence of single family and institutional land uses may be correlated with lower levels and more disconnection of impervious cover in the urban core than other high and more directly connected impervious cover associated with commercial and industrial land uses.

In evaluation of regression results, it is critical to remember that "correlation does not equal causation". Further examination of specific factors would be necessary before attempting to formulate a management strategy purely based on these regression results, particularly in relation to the specific land use factors. For example, in western watersheds the positive relationship with transportation infrastructure and bacteria counts may be reflective of development density (more development, more roads) or increased loading from birds or bats on bridges (Sejkora et al., 2011). The positive relationship for industrial uses may simply be a function of increasing impervious cover and reduced stormwater infiltration. While the inverse relationship for commercial land use is counter-intuitive, the type of commercial uses (retail versus food service, for example) or spatial characteristics may be the underlying factor driving the significance of this factor in the model. Use of more specific geographic information, including location, frequency and magnitude of sanitary sewer overflows or age of wastewater infrastructure, if available and incorporated into these models may provide additional resolution.

<u>Random Forest Analysis</u>
Modeling *E. coli* for all Austin watersheds combined by Random Forest produced a model with adequate fit and strong emphasis on the length of wastewater lines present in a reach, impervious cover in the

riparian zone, and the density of households. While the Random Forest model does not give specific, directional multipliers like a linear regression model, the variable importance factor provides insight into the factors that may correlate with observed pattens in measured *E. coli*.

When analyzed separately, the Random Forest model of the western watersheds emphasized the percentage of impervious cover within the riparian zone and the percentage of woody vegetation in the riparian zone. Reaches with riparian zones containing high impervious cover and low woody vegetation would have a reduced capacity to infiltrate stormwater or reduce stormwater suspended sediment and associated bacteria load to streams.

The Random Forest model for the eastern watersheds emphasized many parameters including the number of flood detention ponds, commercial land use, population density, and the percentage of impervious cover in the riparian zone. The relatively high importance of many variables complicates interpretation of this model. However, this underscores the potential for greater complexity in *E. coli* patterns in the Blackland Prairie Ecoregion.

The urban watersheds emphasized the length and length averaged diameter of the wastewater collection system in addition to the amount of the area in the reach with transportation (streets, parking lots) land use. As the length of wastewater collection lines was the most important variable in the urban model, the older wastewater infrastructure in the urban core may be contributing to *E. coli* in urban streams although not all potential fecal loading sources in the urban environment were included in the model.

Conclusions

The kriged map of median bacteria counts (Figure 1) show higher density of median non-storm *E. coli* in the urban core and northern (or more developed) portions of the eastern watersheds relative to the western watersheds.

By multiple linear regression analysis using all Austin watersheds, flood detention structural controls were inversely related to bacteria counts while impervious cover in the riparian zone, transportation land use, industrial land use, presence of treated wastewater discharges and household density were positively related to bacteria counts. The citywide Random Forest model also selected as important variables household density and riparian zone impervious cover, although the length of wastewater collection lines was also identified as being related to *E. coli*. Different patterns of significant explanatory variables emerge when watersheds are analyzed separately by geographic group: east of IH-35 in the Blackland Prairie Ecoregion, west of IH-35 in the Edwards Plateau, and the downtown urban core.

In the western watersheds, multiple linear regression yielded inverse relationships between *E. coli* and open space, woody vegetation in the riparian zone, bare ground in the riparian zone and commercial land use but direct relationships between *E. coli* and industrial land use, length of wastewater lines and transportation land use. The western Random Forest model partially aligned with the multiple linear regression, and indicated as important variables the percentage of woody vegetation in the riparian zone in addition to the percentage of impervious cover in the riparian zone.

In the eastern watersheds, influential outlier reaches yielded a model including the presence of treated wastewater discharge permits and household density as significant explanatory variables. Removal of these 3 outlier sampling reaches yielded a model with a low r^2 but with an inverse relationship between *E. coli* and percentage of open space and undeveloped land. The eastern Random Forest model indicated as important variable the number of flood detention structural controls, the percentage of commercial land use, population density and the percentage of impervious cover in the riparian zone. This lack of

alignment in the eastern watersheds may be reflective of an increase in the complexity of the relations between land use and demographic patterns and *E. coli* levels in the Blackland Prairie watersheds.

In the urban core watersheds, multiple linear regression yielded an inverse relationship between *E. coli* and undeveloped land use, single family land use, institutional land use and flood detention structural controls but direct relationships with length of wastewater collection lines, sparse grasses in the riparian zone and woody vegetation in the riparian zone. This partially aligned with the Random Forest model which selected as important variables wastewater infrastructure and percent transportation land use.

References

Breiman, L. 2001. Random forests. Machine Learning 45(1): 5-32.

Breiman, L. 2002. Manual on Setting Up, Using, and Understanding Random Forests v3.1. unpublished manuscript, http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf.

City of Austin. 2002. Environmental Integrity Index Methodology. City of Austin Watershed Protection Department, Environmental Resource Management Division. COA-ERM 1999-01. http://assets.austintexas.gov/watershed/publications/files/EII%20May%2031%202002%20edition.pdf

City of Austin. 2013. Austin Watershed Regulation Areas. Map retrieved May 7, 2013, from: ftp://ftp.ci.austin.tx.us/GIS-Data/Regional/standard_maps/Watershed%20Regulation%20Areas.pdf

Crabill, C., and R. Donald, J. Snelling, R. Foust, G. Southam. 1999. The impact of sediment fecal coliform reservoirs on seasonal water quality in Oak Creek, Arizona. Wat. Res. Vol. 33, No. 9, pp. 2163-2171.

Craig, D.L., Fallowfield, H.J., Cromar, N. J. 2004. Use of Microcosms to Etermine Persistence of *Escherichia coli* in Recreational Coastal Water and Sediment and Validation With In Situ Measurements. Journal of Applied Microbiology vol.96, n°6, pp. 922-930.

Downing, J.A. 1979. Aggregation, transformation, and the design of benthic sampling programs. Journal of the Fisheries Research Board of Canada 36: 1454-1463.

Downing, J.A. 1981. How well does the 4th root transformation work? Canadian Journal of Fisheries and Aquatic Science 38: 127-129.

Gromping, U. 2009. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician 63(4): 308-319.

Isaaks, E. H, and R. M. Srivastava. 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York.

Ishwaran, H. 2007. Variable Importance in Binary Regression Trees and Forests. Electronic Journal of Statistics 1: 519-537.

Montgomery, D. C., E. A. Peck, and G. G. Vining. 2006. Introduction to Linear Regression Analysis. Fourth Edition. John Wiley and Sons, Inc. Hoboken, New Jersey.

Richter, A., and A. Duncan. 2012. Riparian Functional Assessment: Choosing Metrics that Quantify Restoration Success in Austin, Texas. City of Austin Watershed Protection Department. SR-12-12.

Scoggins, M., and A. Richter, J. Meade, K. Murray. 2013. An Index of Riparian Integrity for the Austin Area. City of Austin Watershed Protection Department. SR-13-09.

Sejkora, P., and M.J. Kirisits, M. Barrett. 2011. Colonies of Cliff Swallows on Highway Bridges: A Source of *Escherichia coli* in Surface Waters. Journal of the American Water Resources Association 47(6):1275–1284.

Sjogren, R. E. 1994. Prolonged Survival of an Environmental *Escherichia coli* in Laboratory Soil Microcosms. Water, Air and Soil Pollution 75: 389-403.

Sokal, R.R., and F.J. Rohlf. 1995. Biometry: the principles and practice of statistics in biological research. W.H. Freeman and Company, New York.

Stephenson, G.R. and R.C. Rychert. 1982. Bottom Sediment: a Reservoir of Escherichia coli in Rangeland Streams. Journal of Range Management 35(1), January 1982, pp. 119-123.

Strobl, C., A. Boulesteix, A. Zeiles, and T. Hothorn. 2007. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics 8: 25.

Strobl, C., J. Malley, and G. Tutz. 2009. An Introduction to Recursive Partitioning: Rational, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. Psychological Methods 14(4): 323-348.

Texas Commission on Environmental Quality (TCEQ). 2012. Draft 2012 Texas Integrated Report for Clean Water Act sections 305(b) and 303(d). http://www.tceq.texas.gov/waterquality/assessment/waterquality/assessment/12twqi/twqi12. Accessed 05/14/2013.

Tobler, W.R. 1970. A computer movie simulating urban growth in the Detroit Region. Economic Geography 46: 234-240.

Appendix A

Plots of Bacteria Counts by Geographic Regions
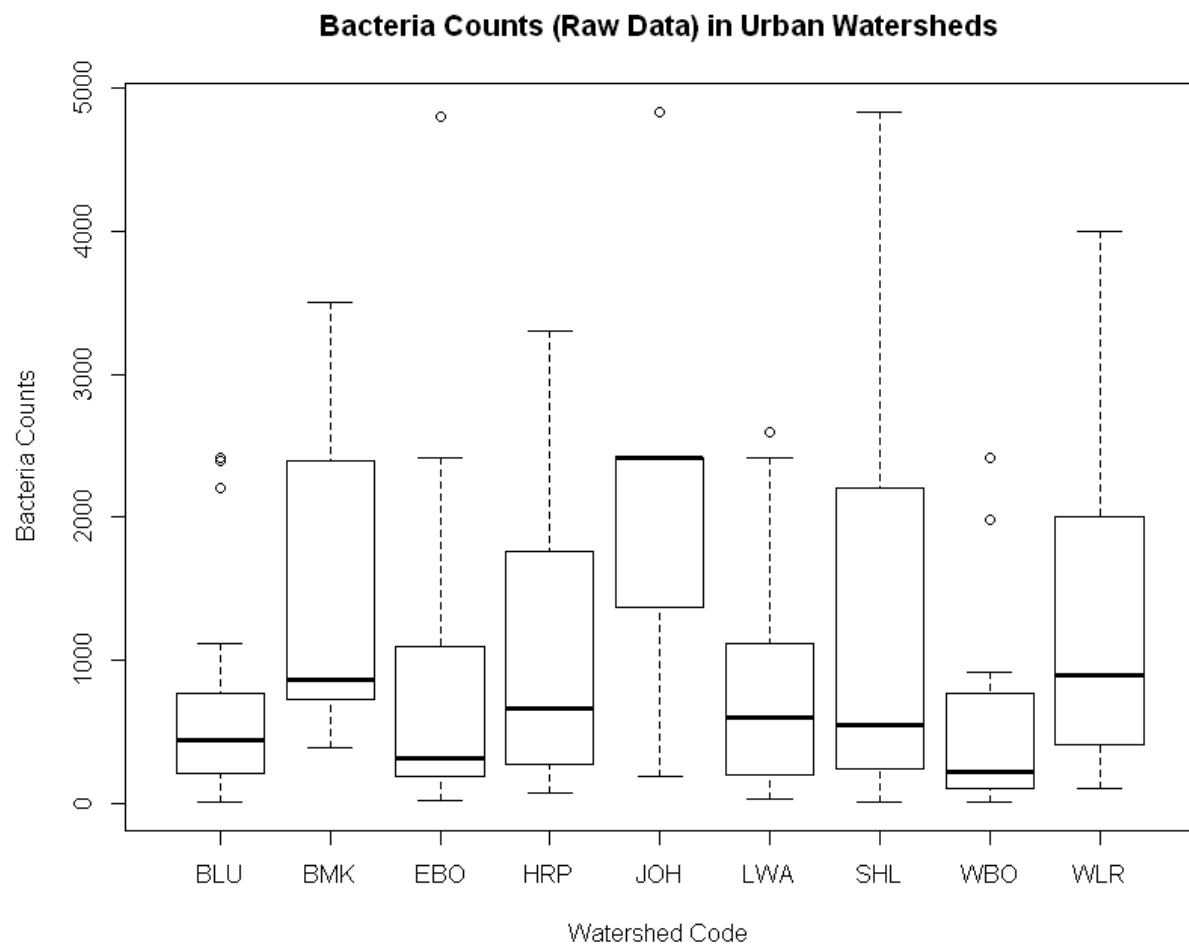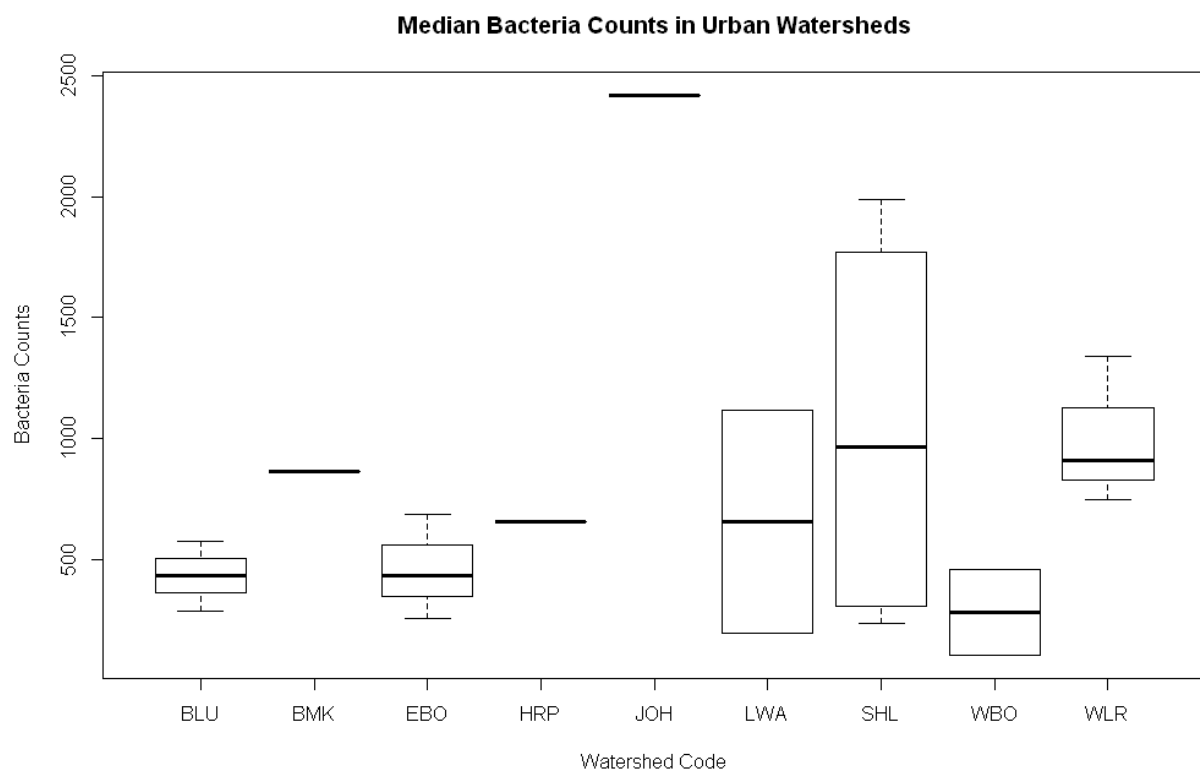
Bacteria Counts (Raw Data) in Urban Watersheds

Median Bacteria Counts in Urban Watersheds

All Bacteria Counts in Eastern Watersheds

Median Bacteria Counts in Eastern Watersheds

Bacteria Counts (Raw Data) in Western Watersheds
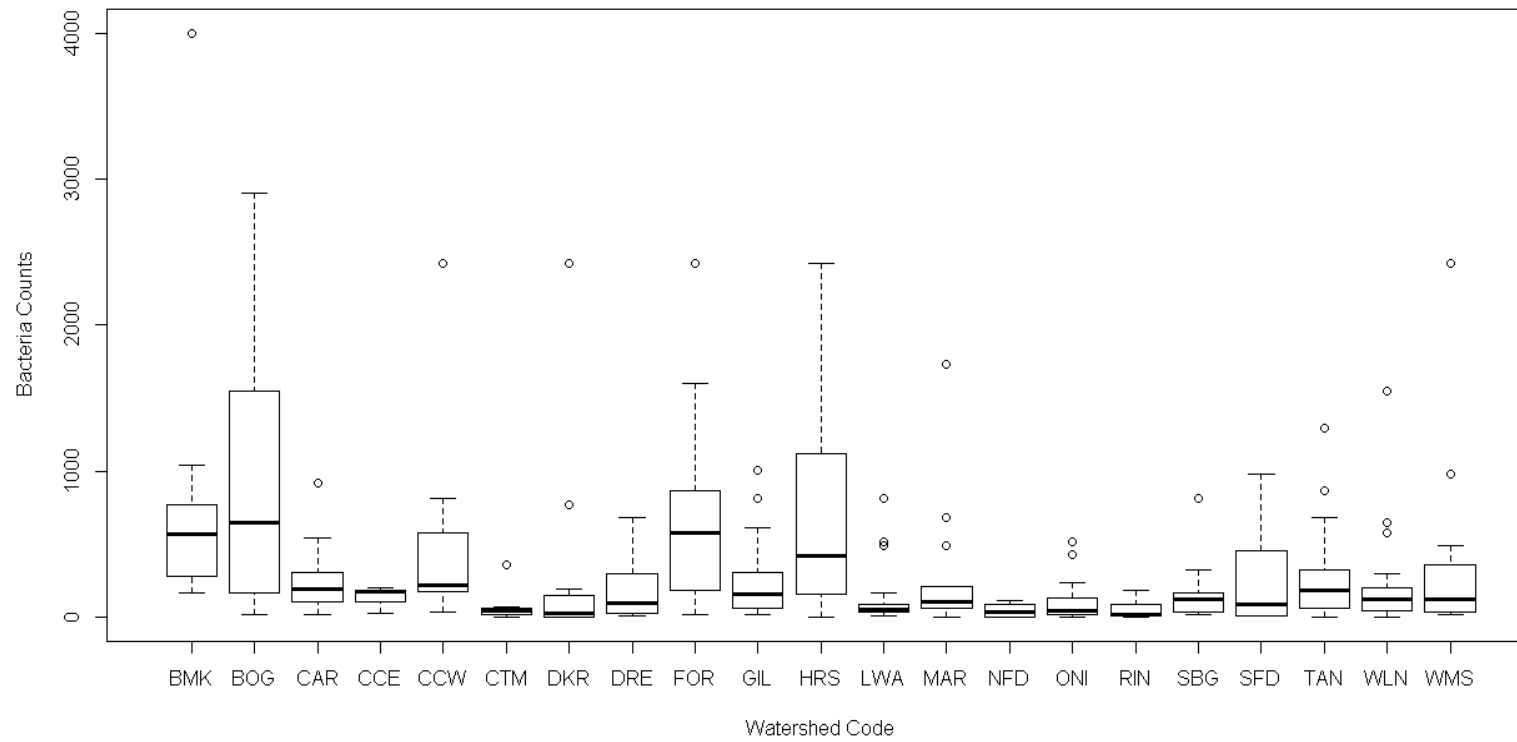
Median Bacteria Counts in Western Watersheds

Appendix B

Tables of Descriptive Statistics

Table B1: Summary Statistics of Bacteria Counts per Watershed Reach

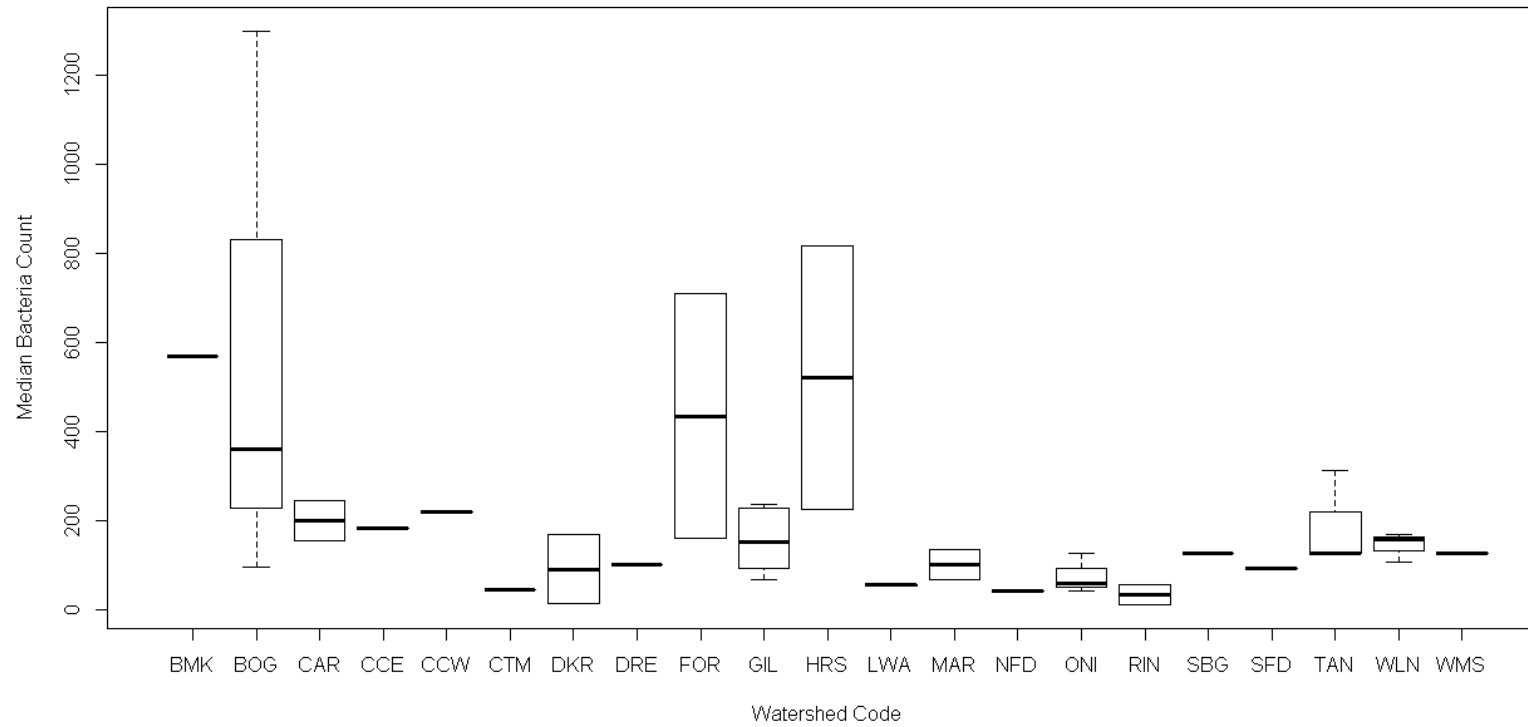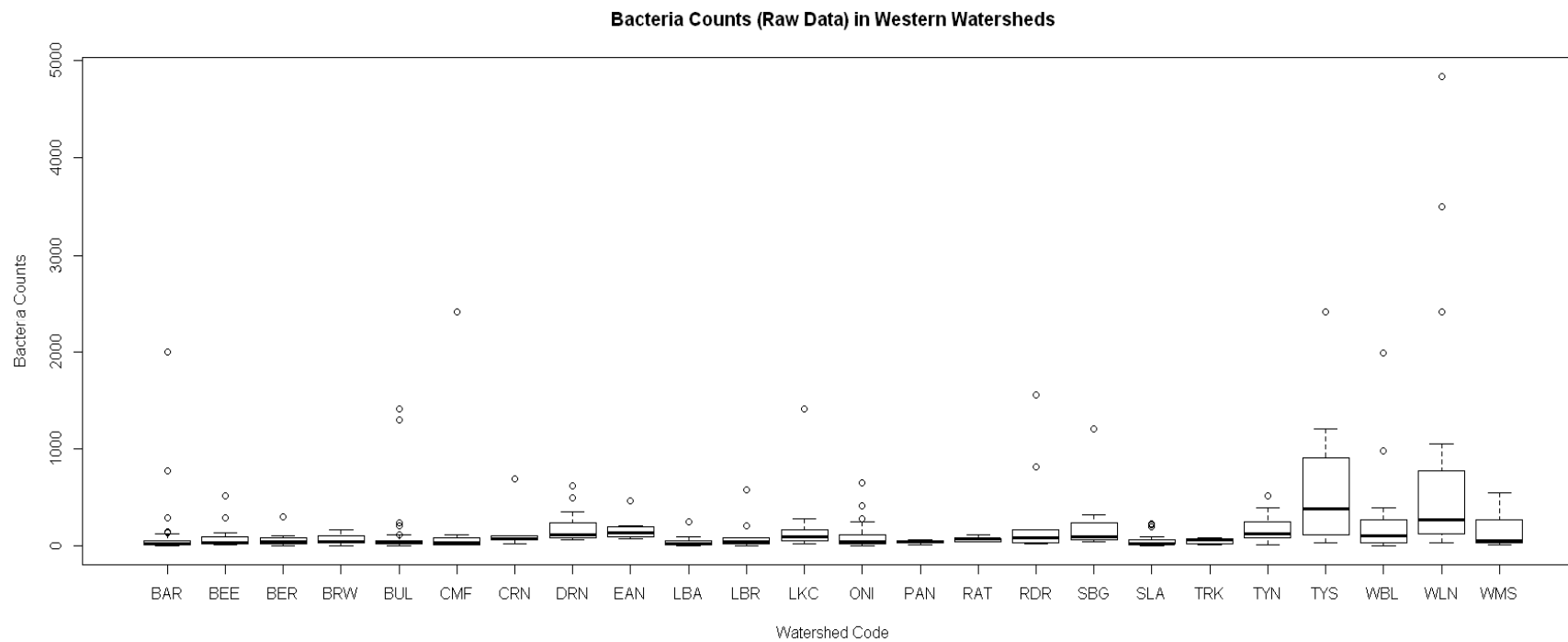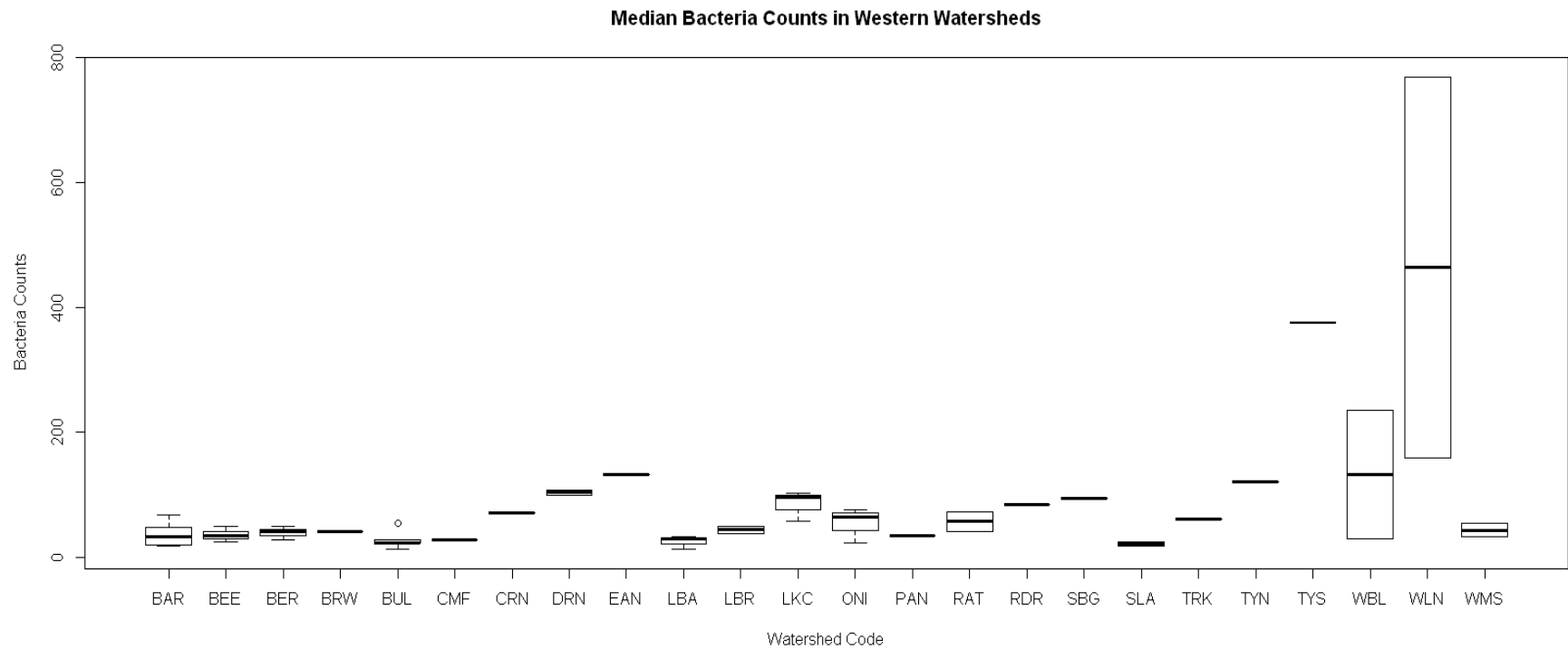| Watershed Reach	N	Mean	Geometric Mean	Median	Standard Deviation	95%-ile	5%-ile
BAR1	6	170.3	45.7	48.0	300.3	615.0	6.5
BAR2	12	30.3	19.2	20.0	37.1	93.1	6.4
BAR3	10	19.9	14.9	18.5	15.2	44.5	4.5
BAR4	13	195.1	43.1	32.0	543.8	884.0	10.2
BAR5	6	43.5	23.0	35.0	39.7	100.0	5.8
BAR6	6	87.2	43.8	68.5	104.2	237.3	7.3
BEE1	10	32.7	25.0	25.6	25.8	74.5	9.9
BEE2	12	122.0	51.2	34.6	160.4	392.7	9.2
BEE3	11	54.6	41.2	50.4	40.5	117.2	12.9
BER1	9	58.8	54.4	50.4	24.3	91.8	33.7
BER2	3	35.7	29.2	41.4	22.4	53.3	14.0
BER3	10	63.6	25.3	27.9	89.4	211.4	2.4
BLU1	10	402.9	194.3	290.0	350.9	885.8	19.5
BLU2	12	922.5	543.7	579.7	904.5	2408.8	102.0
BLU3	11	529.4	321.7	436.0	606.3	1535.0	62.1
BMK1	12	812.2	533.6	568.5	1042.1	2375.4	196.8
BMK3	9	1421.5	1109.9	866.4	1079.7	3067.7	464.1
BOG1	4	93.7	71.5	96.0	60.8	155.7	28.5
BOG2	11	738.0	346.2	360.0	786.1	2076.3	31.5
BOG3	11	1407.8	1079.3	1300.0	916.6	2659.8	333.3
BRW1	11	63.5	32.5	42.0	54.8	147.9	2.1
BUL1	11	316.8	102.9	55.0	519.1	1354.9	21.5
BUL2	10	30.1	20.3	23.4	31.3	79.2	5.8
BUL3	11	18.7	11.3	14.0	22.9	55.0	2.5
BUL4	7	47.9	37.4	23.0	35.2	95.3	18.6
BUL5	7	31.9	27.9	29.0	17.7	56.5	15.3
CAR1	12	317.7	230.6	244.9	295.1	920.8	94.3
CAR2	9	231.8	140.9	155.2	205.7	547.5	26.3
CCE1	3	136.3	99.3	182.0	95.2	198.2	42.4
CCW2	10	518.7	283.6	219.7	708.6	1699.8	71.4
CMF1	8	333.0	40.1	28.4	843.9	1613.7	6.7
CRN1	6	167.7	82.5	71.9	255.8	539.7	26.5
CTM1	7	83.7	32.2	44.8	126.4	277.3	7.2
DKR1	12	40.3	18.0	13.3	58.6	157.0	5.2
DKR3	7	521.7	125.6	167.0	876.8	1924.8	11.4
DRE1	5	719.2	303.5	307.6	767.9	1553.1	49.4
DRE2	8	196.0	92.3	100.4	234.5	581.9	13.8
DRN1	10	129.2	118.1	108.3	61.5	233.3	75.1
DRN2	10	219.3	149.4	100.6	202.2	557.0	59.8

Watershed Reach	N	Mean	Geometric Mean	Median	Standard Deviation	95%-ile	5%-ile
EAN2	9	166.7	141.1	133.4	120.0	360.6	76.7
EBO1	5	552.1	434.7	435.2	400.6	1045.9	193.7
EBO2	11	1337.2	457.4	686.7	1497.7	3609.8	27.3
EBO3	5	198.4	162.2	260.3	118.3	304.2	68.8
FOR3	3	112.3	77.1	159.0	81.7	159.9	32.1
FOR4	10	904.6	705.0	710.0	678.6	2050.6	240.2
GIL1	12	159.7	118.2	125.6	116.0	353.7	30.0
GIL2	11	202.4	98.4	66.3	265.8	714.8	20.9
GIL3	11	126.9	96.4	93.2	93.1	271.8	29.6
GIL4	8	258.1	121.9	179.8	329.0	778.0	21.3
GIL5	12	229.8	177.6	226.5	140.9	404.0	50.7
GIL6	12	282.8	216.8	236.4	191.9	594.6	64.6
HRP1	8	1095.8	562.5	658.8	1163.9	2991.9	75.0
HRS1	11	433.2	168.7	224.7	681.4	1454.0	13.7
HRS2	11	994.2	631.9	816.4	784.7	2203.0	90.1
JOH1	7	2146.7	1388.4	2419.6	1566.1	4113.3	226.6
LBA1	11	17.7	11.1	13.4	13.7	37.6	1.5
LBA2	10	60.5	38.9	33.8	71.2	177.3	12.5
LBA3	9	41.0	24.4	31.1	33.5	81.4	4.1
LBR1	9	74.3	46.1	38.9	77.1	206.0	11.6
LBR2	7	114.1	29.4	50.4	207.4	430.0	3.2
LKC1	9	128.3	106.1	95.9	80.3	243.7	47.3
LKC2	9	261.4	135.1	103.9	437.2	958.4	50.1
LKC3	11	97.6	71.7	58.1	77.8	229.2	22.3
LWA1	12	86.8	51.8	56.5	129.0	272.6	13.8
LWA2	12	164.3	80.2	52.8	246.6	651.8	24.8
LWA3	11	275.4	161.2	200.0	238.5	604.7	33.7
LWA4	11	1399.5	1194.2	1119.9	776.0	2509.8	499.4
MAR1	10	360.7	150.3	133.5	528.1	1262.1	18.4
MAR2	3	47.8	24.5	67.0	38.9	72.7	9.4
NFD1	4	50.9	24.7	40.8	53.7	110.7	5.1
ONI1	11	83.2	51.9	59.0	77.7	210.2	13.0
ONI2	11	54.1	31.7	41.0	66.8	163.0	9.3
ONI3	11	153.6	92.2	127.0	168.1	476.0	27.4
ONI4	11	107.6	68.2	77.1	93.8	262.0	15.4
ONI5	10	65.8	53.2	65.0	40.7	125.2	17.6
ONI6	10	120.1	24.8	23.2	223.8	542.5	3.7
PAN1	7	38.8	35.2	35.9	16.3	59.0	17.3
RAT1	3	64.7	57.8	42.6	39.7	103.8	41.2
RDR1	9	314.6	97.5	85.5	528.8	1258.4	19.4
RIN1	11	63.3	41.4	56.5	55.2	143.4	11.0

Watershed Reach	N	Mean	Geometric Mean	Median	Standard Deviation	95%-ile	5%-ile
RIN2	3	8.2	7.6	9.7	3.6	10.8	4.7
SBG1	9	190.7	99.5	125.9	253.3	620.0	23.2
SBG2	8	253.5	128.1	95.2	393.9	891.7	42.3
SFD1	10	235.8	88.5	93.2	320.0	772.0	13.3
SHL1	13	1831.9	1385.2	1553.1	1215.5	3387.6	356.6
SHL2	13	1690.6	865.6	1986.3	1423.9	3600.0	100.4
SHL3	11	329.1	167.0	240.0	325.7	908.2	22.7
SHL4	10	562.3	290.9	377.1	557.7	1448.1	35.2
SLA1	11	49.3	28.2	19.0	62.6	157.3	8.6
SLA3	9	67.0	28.6	26.0	85.1	216.5	6.4
TAN1	7	148.7	72.0	127.4	128.4	315.1	6.7
TAN2	10	204.5	118.9	122.0	253.2	623.0	30.0
TAN3	9	453.1	289.3	312.3	387.0	1054.5	82.1
TRK1	5	50.2	38.4	62.3	32.2	80.3	12.3
TYN1	10	181.6	122.6	121.8	157.7	456.2	31.2
TYS1	12	605.8	318.7	376.4	689.0	1750.6	52.0
WBL1	9	489.9	268.4	235.9	624.9	1583.9	64.0
WBL2	10	67.1	28.1	30.9	101.0	232.2	5.0
WBO2	11	912.7	481.7	461.1	914.8	2419.6	99.0
WBO3	10	193.9	112.3	106.1	262.6	605.4	30.8
WLN1	9	114.3	92.1	107.0	80.2	243.2	36.6
WLN2	12	268.0	73.1	156.5	440.1	1055.6	2.1
WLN3	11	185.5	127.6	167.0	156.9	441.0	22.5
WLN4	13	1256.2	725.7	770.0	1441.6	4035.7	140.6
WLN5	13	243.0	156.4	160.0	277.5	748.5	51.3
WLR1	14	1262.9	835.4	913.5	1116.6	2972.7	205.6
WLR2	12	969.9	734.0	748.6	752.4	2419.6	243.5
WLR3	8	1435.6	1021.4	1340.2	927.5	2419.6	267.8
WMS1	12	396.1	138.5	126.7	694.0	1628.0	23.6
WMS2	5	215.0	73.8	33.0	266.0	530.2	14.5
WMS3	4	103.5	75.0	55.5	106.9	233.2	41.1

Table B2: Table of Summary Statistics of Transformed Bacteria Counts per Watershed Reach

Watershed Reach	N	Mean	Geometric Mean	Median	Standard Deviation
BAR1	6	2.85	2.60	2.54	1.39
BAR2	12	2.15	2.09	2.11	0.54
BAR3	10	2.00	1.97	2.07	0.41
BAR4	13	2.75	2.56	2.38	1.29
BAR5	6	2.33	2.19	2.42	0.76
BAR6	6	2.71	2.57	2.88	0.92
BEE1	10	2.27	2.24	2.25	0.45
BEE2	12	2.84	2.68	2.41	1.04
BEE3	11	2.58	2.53	2.66	0.52
BER1	9	2.73	2.72	2.66	0.29
BER2	3	2.36	2.32	2.54	0.47
BER3	10	2.41	2.24	2.29	0.92
BLU1	10	4.00	3.73	4.10	1.36
BLU2	12	5.01	4.83	4.91	1.41
BLU3	11	4.38	4.24	4.57	1.18
BMK1	12	4.92	4.81	4.85	1.17
BMK3	9	5.86	5.77	5.43	1.11
BOG1	4	2.97	2.91	3.13	0.66
BOG2	11	4.58	4.31	4.36	1.58
BOG3	11	5.85	5.73	6.00	1.16
BRW1	11	2.54	2.39	2.55	0.82
BUL1	11	3.01	2.83	2.44	1.24
BUL2	10	2.18	2.12	2.20	0.52
BUL3	11	1.89	1.84	1.93	0.51
BUL4	7	2.51	2.47	2.19	0.48
BUL5	7	2.32	2.30	2.32	0.33
CAR1	12	3.97	3.90	3.94	0.84
CAR2	9	3.58	3.45	3.53	1.00
CCE1	3	3.24	3.16	3.67	0.83
CCW2	10	4.26	4.10	3.85	1.26
CMF1	8	2.83	2.52	2.29	1.77
CRN1	6	3.14	3.01	2.91	1.05
CTM1	7	2.57	2.38	2.59	1.01
DKR1	12	2.16	2.06	1.90	0.75
DKR3	7	3.75	3.35	3.59	1.88
DRE1	5	4.49	4.17	4.19	1.79

Watershed Reach	N	Mean	Geometric Mean	Median	Standard Deviation
DRE2	8	3.28	3.10	3.16	1.14
DRN1	10	3.31	3.30	3.23	0.37
DRN2	10	3.58	3.50	3.15	0.85
EAN2	9	3.48	3.45	3.40	0.53
EBO1	5	4.64	4.57	4.57	0.92
EBO2	11	5.10	4.62	5.12	2.14
EBO3	5	3.62	3.57	4.02	0.66
FOR3	3	3.06	2.96	3.55	0.86
FOR4	10	5.24	5.15	5.15	0.99
GIL1	12	3.37	3.30	3.35	0.71
GIL2	11	3.30	3.15	2.85	1.09
GIL3	11	3.19	3.13	3.11	0.64
GIL4	8	3.50	3.32	3.58	1.22
GIL5	12	3.72	3.65	3.88	0.73
GIL6	12	3.91	3.84	3.92	0.77
HRP1	8	5.13	4.87	5.05	1.68
HRS1	11	3.87	3.60	3.87	1.47
HRS2	11	5.20	5.01	5.35	1.37
JOH1	7	6.33	6.10	7.01	1.70
LBA1	11	1.90	1.83	1.91	0.51
LBA2	10	2.56	2.50	2.41	0.64
LBA3	9	2.32	2.22	2.36	0.67
LBR1	9	2.69	2.61	2.50	0.72
LBR2	7	2.58	2.33	2.66	1.24
LKC1	9	3.25	3.21	3.13	0.54
LKC2	9	3.53	3.41	3.19	1.08
LKC3	11	2.97	2.91	2.76	0.63
LWA1	12	2.76	2.68	2.74	0.73
LWA2	12	3.11	2.99	2.69	1.00
LWA3	11	3.71	3.56	3.76	1.07
LWA4	11	5.94	5.88	5.78	0.89
MAR1	10	3.72	3.50	3.39	1.37
MAR2	3	2.37	2.23	2.86	0.91
NFD1	4	2.36	2.23	2.36	0.88
ONI1	11	2.77	2.68	2.77	0.74
ONI2	11	2.45	2.37	2.53	0.68
ONI3	11	3.20	3.10	3.36	0.88
ONI4	11	2.97	2.87	2.96	0.78
ONI5	10	2.74	2.70	2.83	0.48
ONI6	10	2.27	2.11	2.12	0.95
PAN1	7	2.45	2.44	2.45	0.30

| Watershed Reach	N	Mean	Geometric Mean	Median	Standard Deviation
RAT1	3	2.78	2.76	2.55	0.40
RDR1	9	3.38	3.14	3.04	1.48
RIN1	11	2.61	2.54	2.74	0.66
RIN2	3	2.56	2.48	2.50	0.77
SBG1	9	3.29	3.16	3.35	1.03
SBG2	8	3.49	3.36	3.12	1.11
SFD1	10	3.29	3.07	3.11	1.31
SHL1	13	6.23	6.10	6.28	1.26
SHL2	13	5.75	5.42	6.68	1.86
SHL3	11	3.81	3.59	3.94	1.24
SHL4	10	4.36	4.13	4.40	1.40
SLA1	11	2.38	2.31	2.09	0.68
SLA3	9	2.47	2.31	2.26	0.91
TAN1	7	3.11	2.91	3.36	1.09
TAN2	10	3.42	3.30	3.31	0.96
TAN3	9	4.28	4.12	4.20	1.12
TRK1	5	2.54	2.49	2.81	0.54
TYN1	10	3.42	3.33	3.32	0.83
TYS1	12	4.43	4.23	4.40	1.38
WBL1	9	4.21	4.05	3.92	1.26
WBL2	10	2.45	2.30	2.36	0.89
WBO2	11	4.93	4.68	4.63	1.54
WBO3	10	3.37	3.26	3.21	0.94
WLN1	9	3.14	3.10	3.22	0.56
WLN2	12	3.27	2.92	3.54	1.48
WLN3	11	3.46	3.36	3.59	0.81
WLN4	13	5.38	5.19	5.27	1.52
WLN5	13	3.63	3.54	3.56	0.90
WLR1	14	5.53	5.38	5.48	1.36
WLR2	12	5.30	5.21	5.23	1.06
WLR3	8	5.82	5.65	6.01	1.35
WMS1	12	3.66	3.43	3.35	1.47
WMS2	5	3.18	2.93	2.40	1.43
WMS3	4	3.00	2.94	2.72	0.70

Appendix C

**Diagnostic Plots
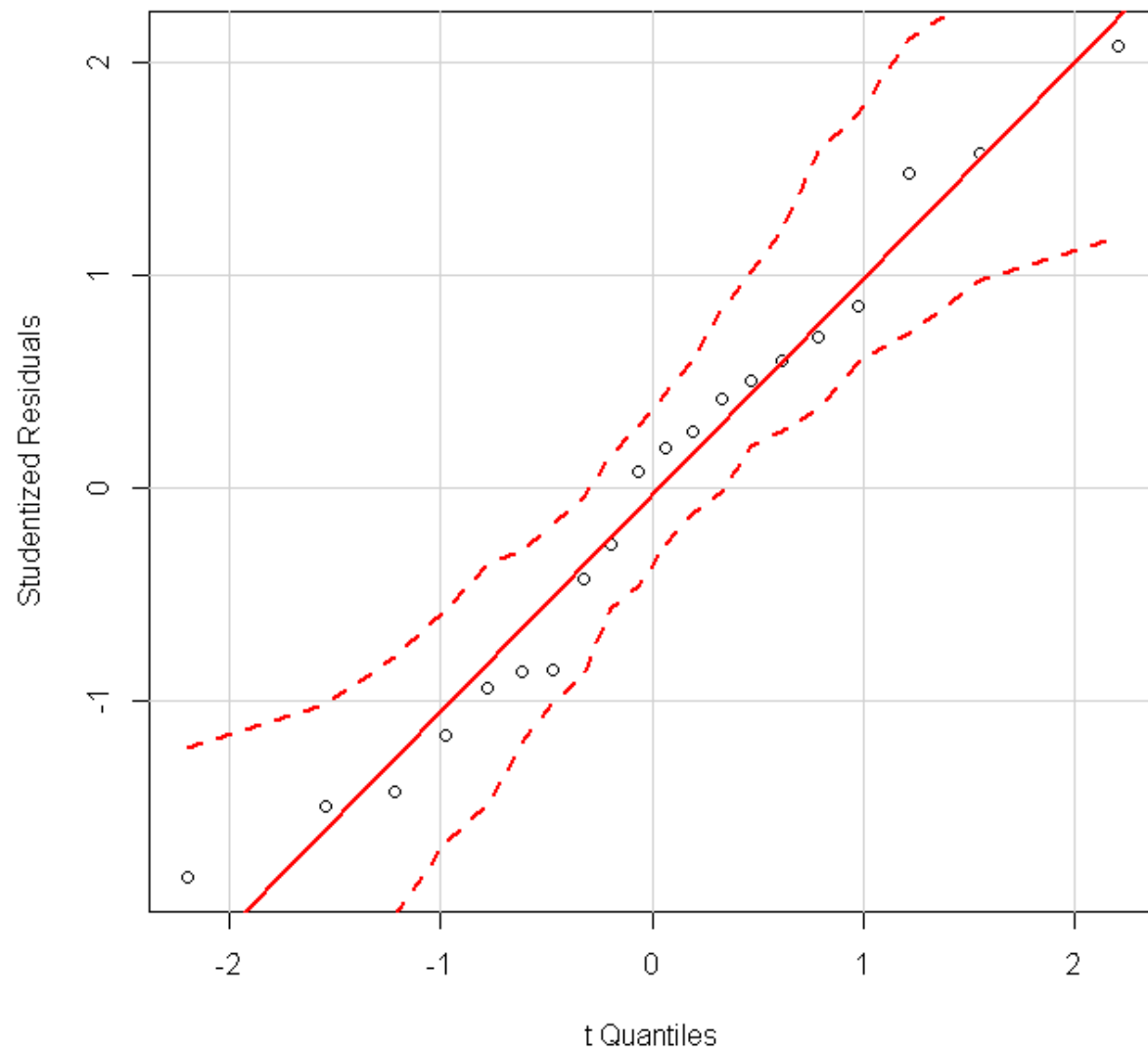and Bacteria Counts for each Watershed**

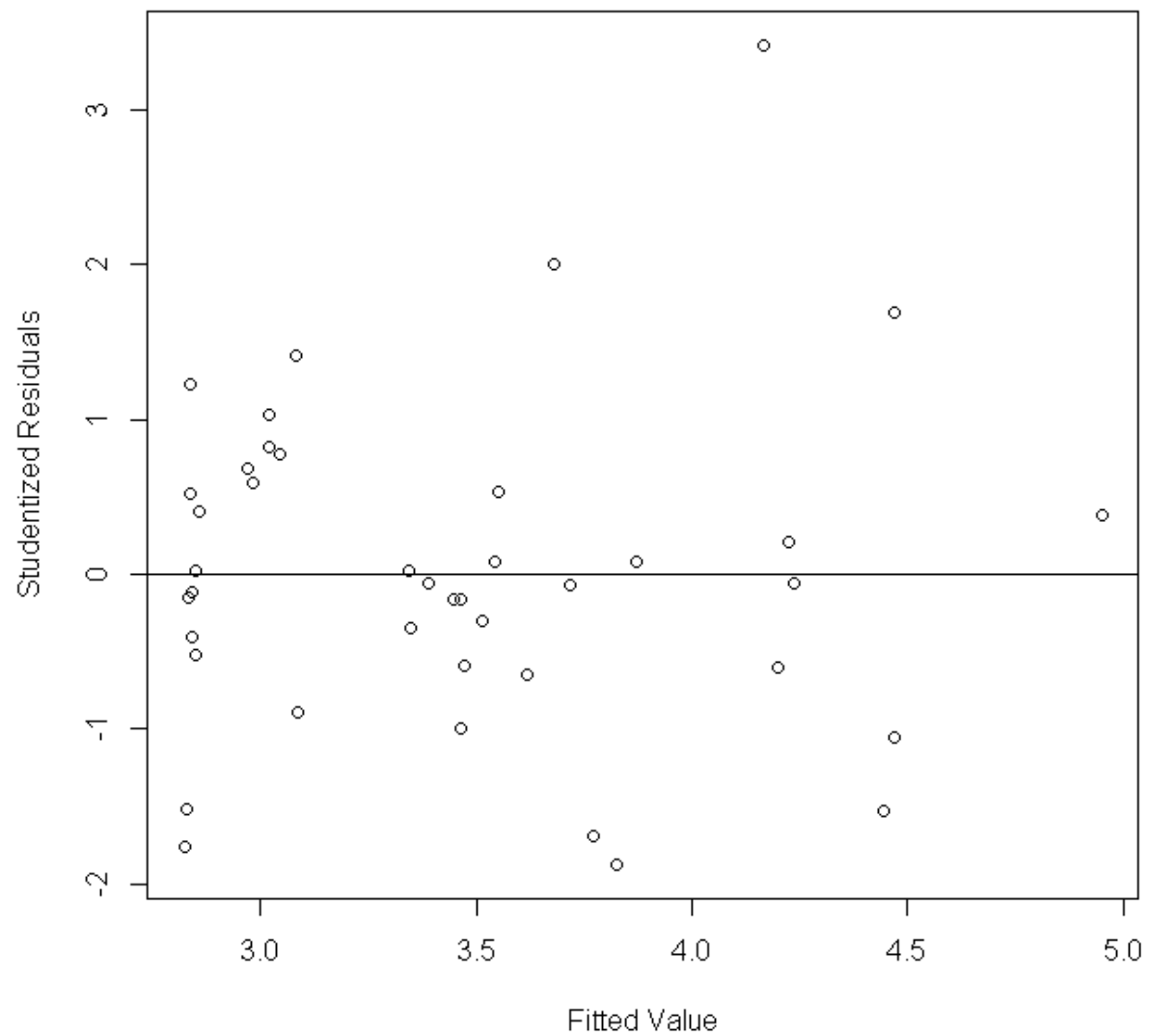# Studentized Residuals of Urban Watersheds
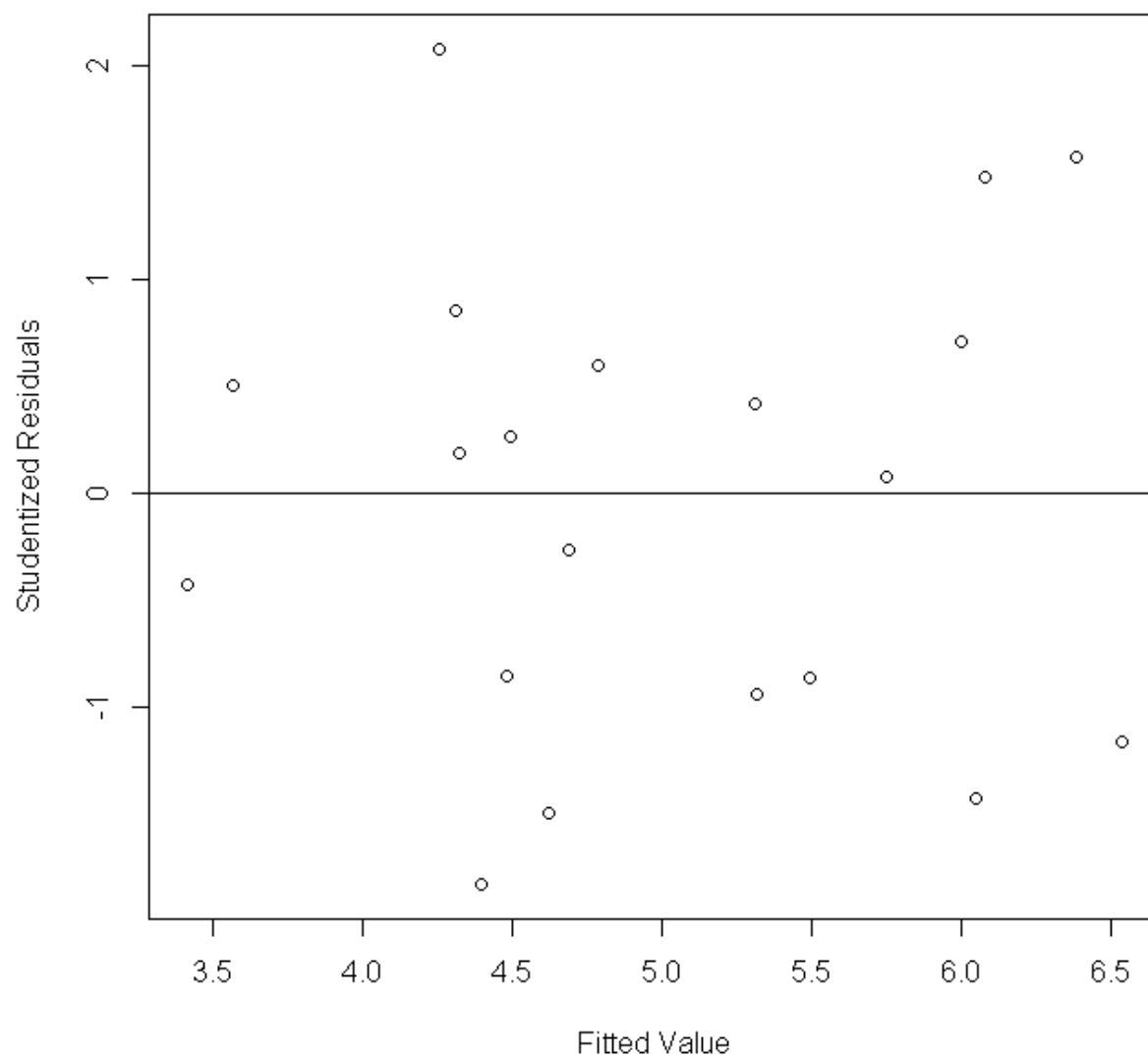
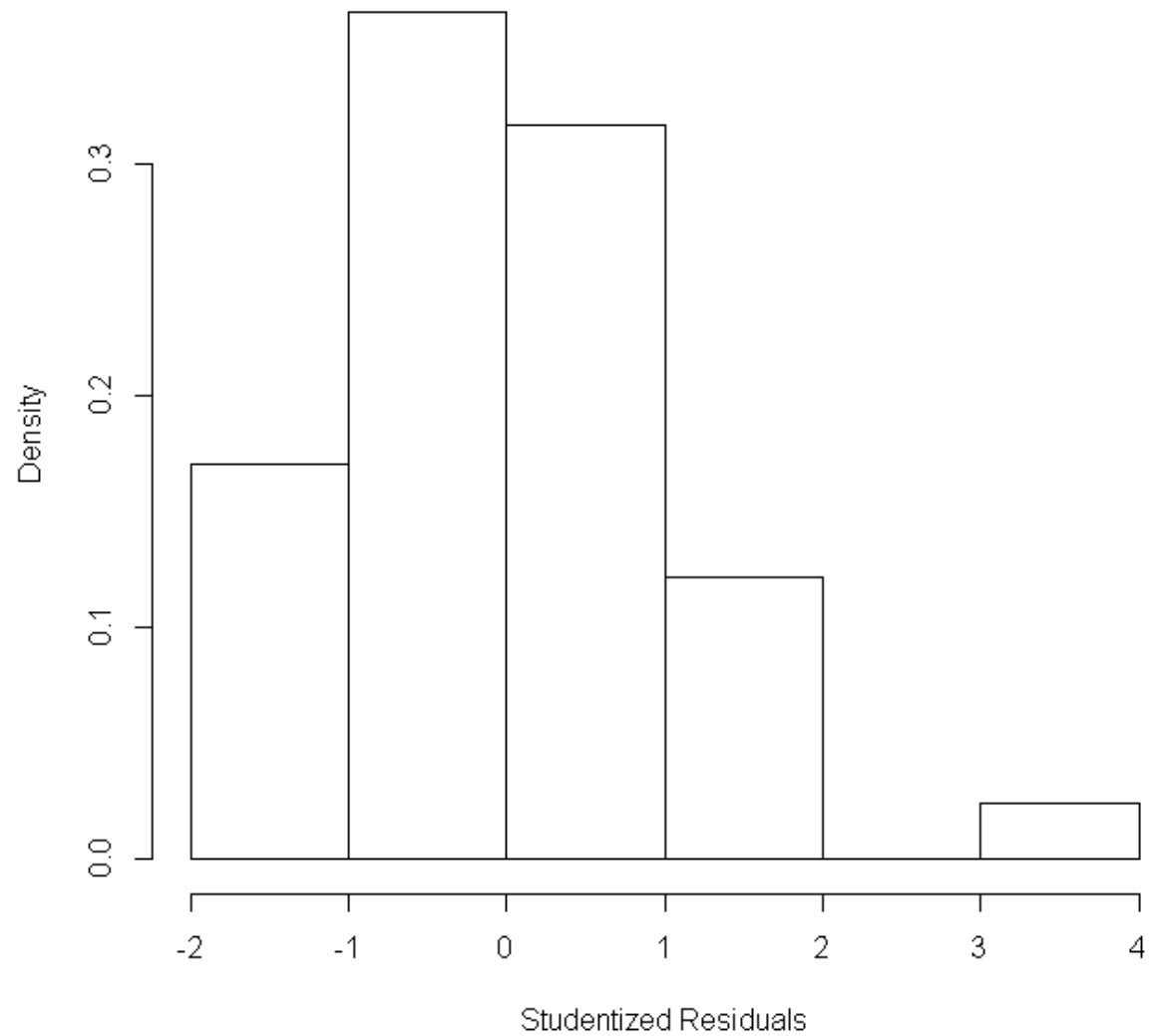# Histogram of Studentized Residuals for Urban Watersheds
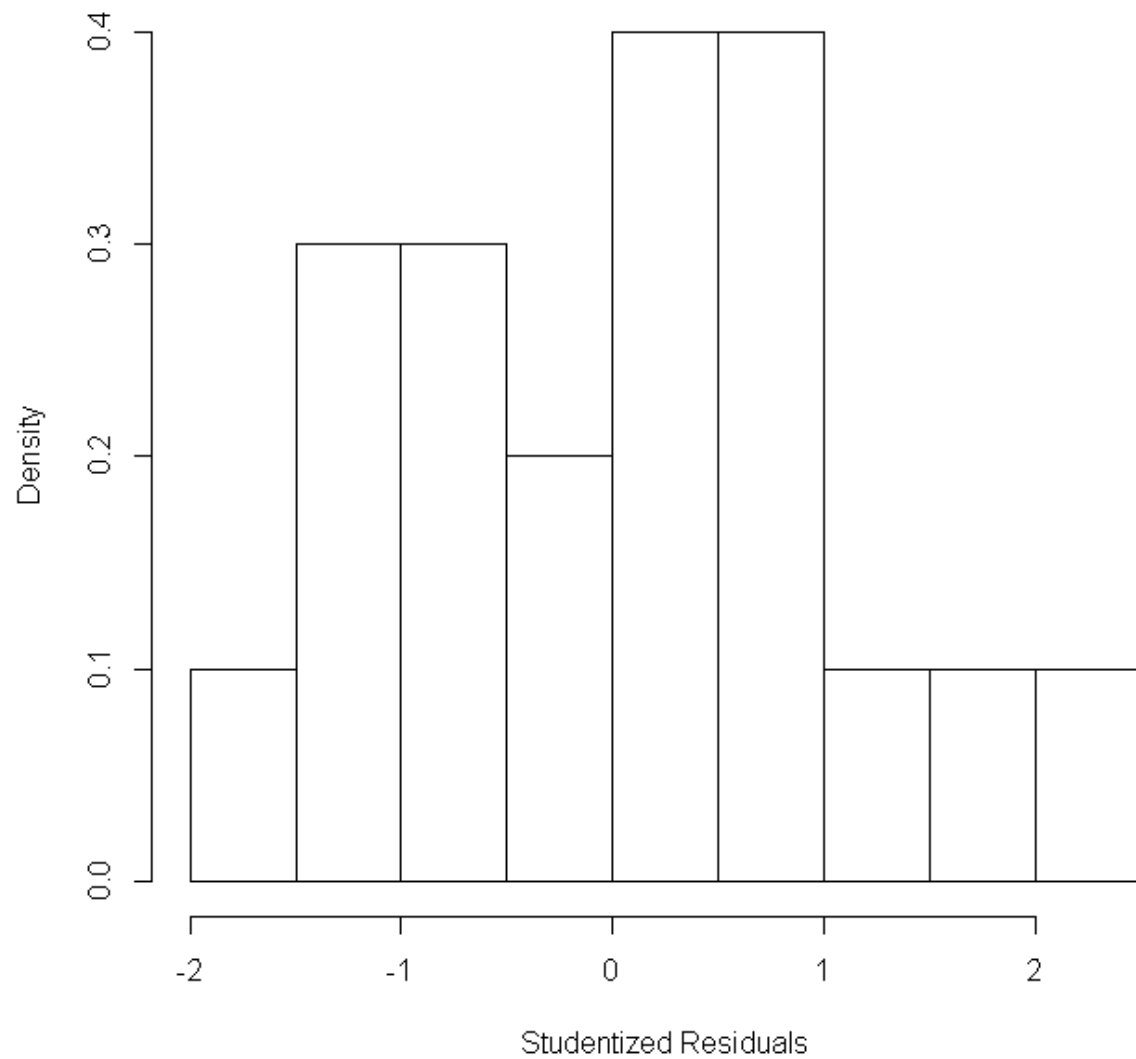
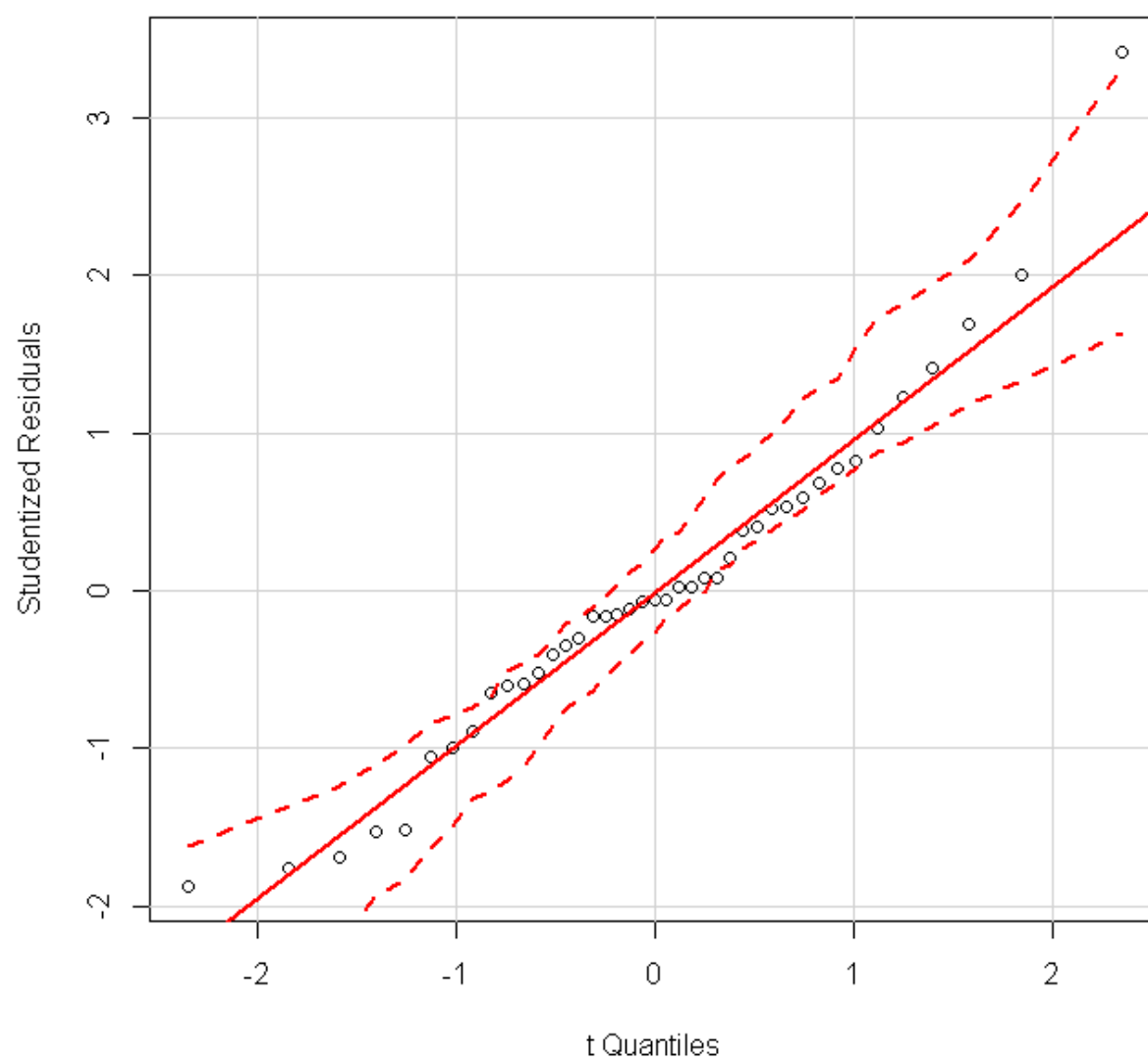## Normal QQ Plot for Urban Watersheds
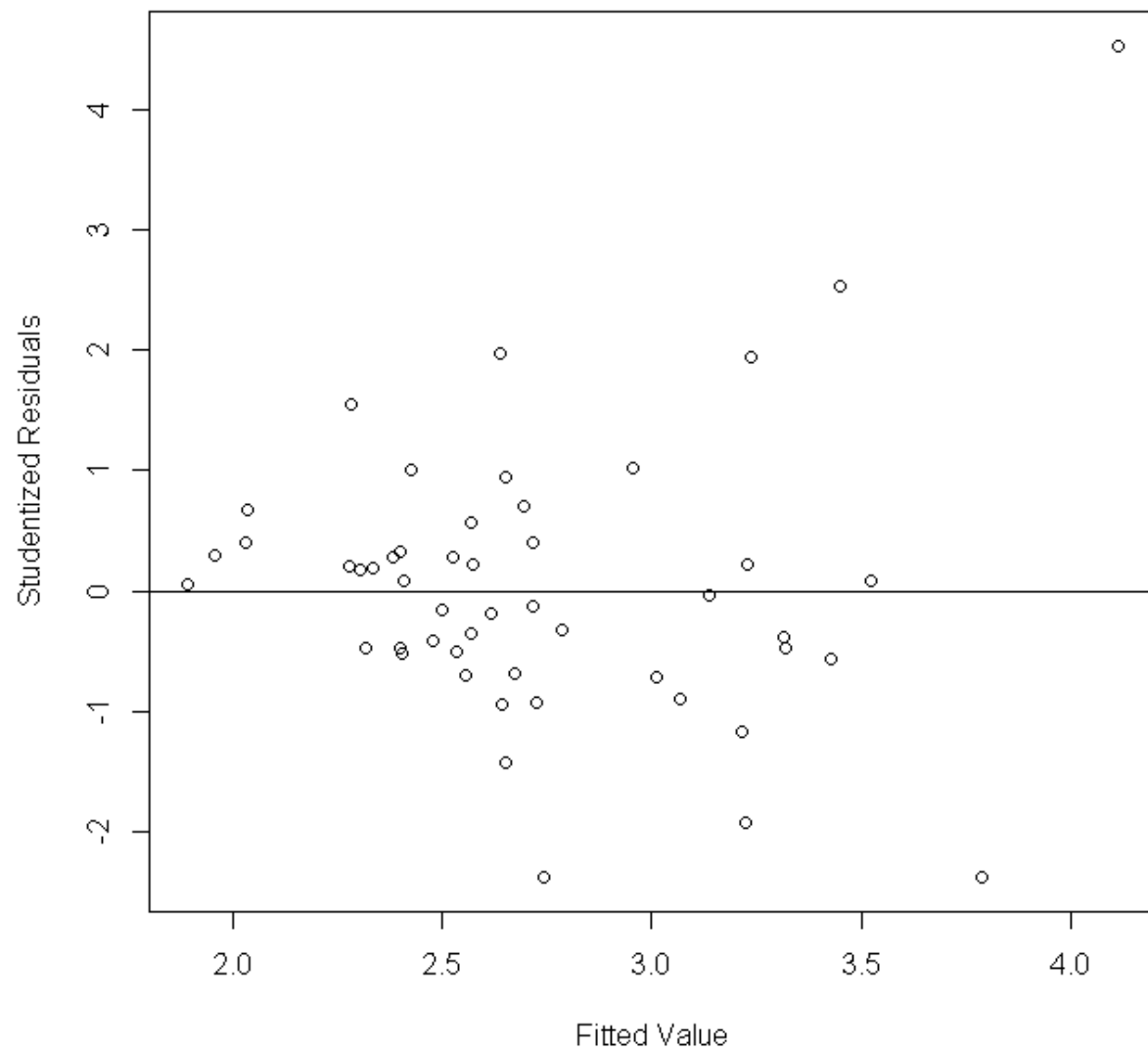
Studentized Residuals of Eastern Watersheds

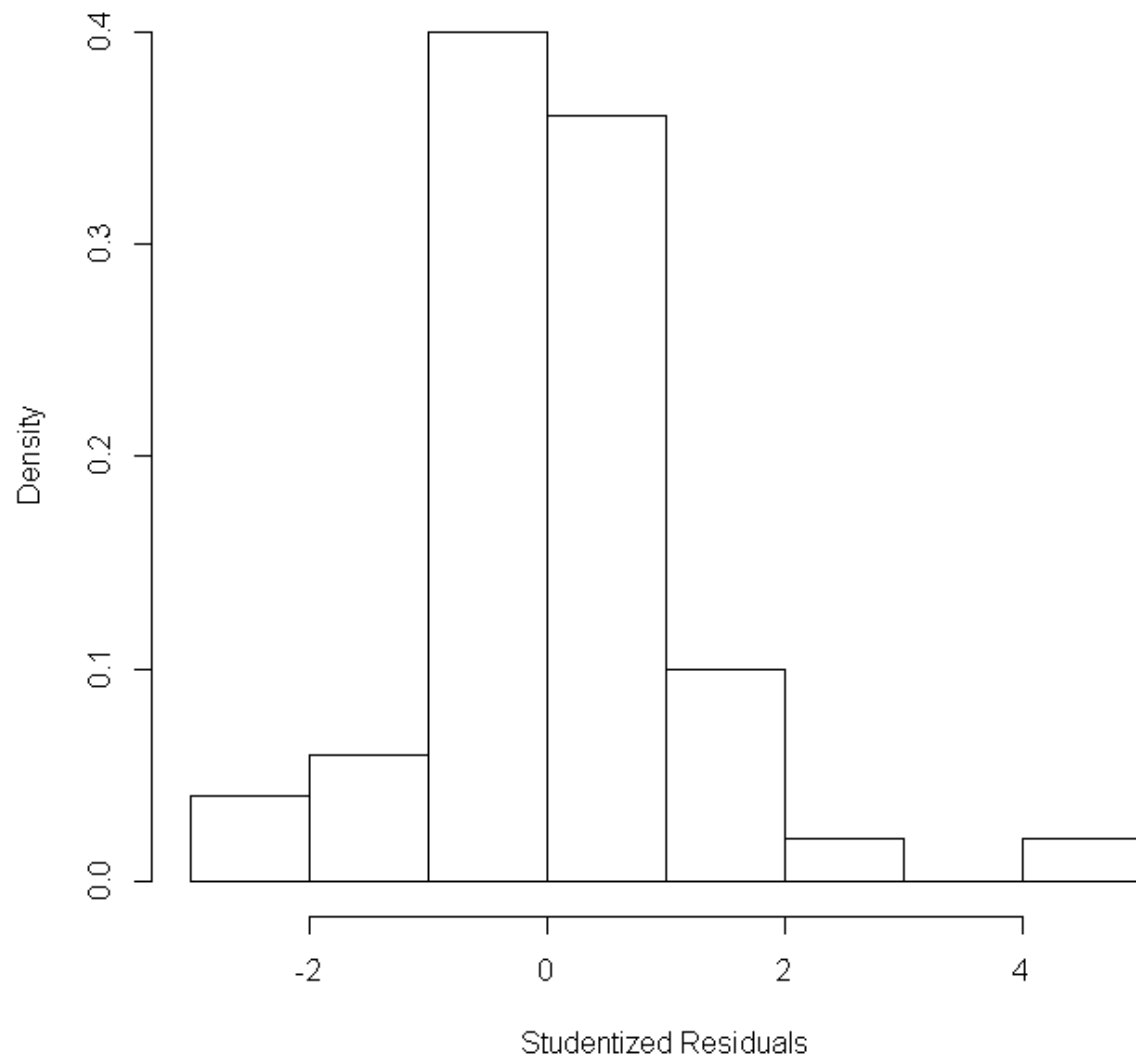# Histogram of Studentized Residuals for Eastern Watersheds

Normal QQ Plot for Eastern Watersheds

Studentized Residuals of Western Watersheds

Histogram of Studentized Residuals for Western Watersheds

Normal QQ Plot for Western Watersheds